Experimental and Quasi-Experimental Designs

For Generalized Causal Inferences

William R. Shadish

Thomas D. Cook

Donald T. Campbell

طرحهای آزمایشی و شبه‌آزمایشی

برای استنباطهای علّی تعمیم‌یافته

ویلیام شَدیش ، توماس کوک و دونالد کمپبل

ترجمه:

آزاده کاظمی نیا

(عضو هیات علمی دانشگاه گیلان)

پیشگفتار مترجم

کتابی که پیش رو دارید به نحوه انجام طرحهای پژوهشی آزمایشی و شبه‌آزمایشی در مطالعات علوم اجتماعی می‌پردازد. این مجموعه به شیوه‌ی رایج کتابهای آماری نوشته نشده و به ندرت، و تنها در موارد بسیار ضروری به بیان مطالب مربوط به محاسبات آماری می‌پردازد. در عوض، تمرکز کتاب بر ملاحضات مرتبط با بکارگیری عناصر طرحهای آزمایشی به منظور بهبود گزاره‌های علّی و تعمیم‌پذیری آنهاست. در طول فصول مختلف کتاب، انواع ابزارهای طرحهای آزمایشی و شبه‌آزمایشی که بکارگیری آنها می‌تواند روایی گزاره های علّی در مطالعات حوزه علوم اجتماعی را افزایش دهد، معرفی شده و مورد بحث قرار می‌گیرند. بخش قابل‌توجهی از مطالب کتاب به معرفی تهدیدهای مترتب بر روایی در انواع مختلف طرحهای مذکور اختصاص دارد.

کتاب حاضر به عنوان بخشی از سرفصل درس روش تحقیق در مقطع دکتری برای دانشجویان رشته‌های مدیریت و دیگر رشته‌های علوم اجتماعی در دانشگاههای معتبر دنیا مورد تدریس قرار می‌گیرد. امیدوارم ترجمه حاضر بتواند تامین‌کننده نظر اساتید ارجمند روش تحقیق بوده و مطالب این کتاب در سرفصل آموزش روش تحقیق پیشرفته قرار گیرد، و بتواند زمینه‌ساز آشنایی بیشتر محققین و دانشجویان دکتری با شیوه‌های صحیح بکارگیری روشهای آزمایشی به منظور دستیابی به روابط علّی متقن در تحقیقات مدیریت و علوم اجتماعی باشد.

تلاش کرده‌ام تا در حد بضاعت خود ترجمه‌ای روان و نزدیک به متن اصلی در اختیار خوانندگان قرار دهم، اما متن حاضر بی‌تردید دچارکاستی‌های فراوانیست. پیشنهادات و رهنمودهای خوانندگان گرامی کمکی ارزنده در بهبود آن خواهد بود.

بر خود لازم می‌دانم از زحمات استاد ارجمند جناب اقای دکتر مهدی محمودی که با پیشنهادات ارزشمند خود کمک فراوانی به این ترجمه نمودند قدردانی نمایم. همچنین از همراهی استاد ارجمند جناب آقای دکتر محسن اکبری، و مساعدت آقای دکتر ..... و خانم سعیدلو در موسسه مطالعات و پژوهشهای بازرگانی قدردانی و تشکر می‌نمایم. این ترجمه بدون همراهی ارزشمند ایشان میسر نبود.

آزاده کاظمی‌نیا

پیشگفتار
۱. آزمایشها و استنباط علی تعمیم یافته
آزمایش کردن و علّیت
تعریف علّت، اثر و رابطه علّی
علّیت، همبستگی و مزاحم ها
علّتهای قابل دستکاری و غیر قابل دستکاری
توضیح علّی و تفسیر علّی
تفسیر مدرن از آزمایش
آزمایشات تصادفی
شبهآزمایش
آزمایش طبیعی
طرح های غیر آزمایشی
آزمایشها و تعمیم روابط علّی
غالب آزمایشات به صورت کاملا محدود انجام می شوند اما اهداف کلی دارند
روایی سازه: تعمیم علّی به منظور نماینده گی کردن
روایی بیرونی: تعمیم علّی برای چندوجهی سازی
رویکردهایی برای انجام تعمیم های علّی
آزمایشها و فراعلم
نقد کوهانی
نقدهای روانشناسی اجتماعی مدرن
علم و اعتماد
کاربردهایی برای آزمایشها
جهانی بدون علّیت و آزمایشها؟

۲. روایی نتایج آماری و روایی درونی
روایی
یک دسته بندی (نوع شناسی) برای روایی

تهدیدات مترتب بر روایی

روایی نتایج آماری

گزارش نتایج آزمونهای آماری برای هم-تغییری

تهدیدات مترتب بر روایی نتایج آماری

مساله پذیرش فرض صفر

روایی درونی

تهدیدات مترتب بر روایی درونی

تخمین روایی درونی در آزمایهای تصادفی و شبه-آزمایشها

رابطه میان روایی درونی و روایی نتایج آماری

۳. روایی سازه و روایی بیرونی

روایی سازه

چرا استنباطها در مورد سازه ها محل اشکال است

ارزیابی خصوصیات نمونه گیری

تهدیدات مترتب بر روایی سازه

روایی سازه، مناسب سازی پیش-آزمایشی و تعیین مختصات پس-آزمایشی

روایی بیرونی

تهدیدات مترتب بر روایی بیرونی

ثبات اندازه اثر در مقابل ثبات جهت اثر علّی

نمونه گیری تصادفی و روایی بیرونی

نمونه گیری هدفمند و روایی بیرونی

ملحقاتی در باب روابط، تعادل ها و اولویتها

رابطه میان روایی سازه و روایی بیرونی

رابطه میان روایی دورنی و روایی سازه

تعادل ها و اولویتها

خلاصه

۴. طرحهای شبه آزمایشی که یا فاقد گروه کنترل هستند و یا فاقد مشاهداتی در مورد نتایج

منطق شبه-آزمایش به طور خلاصه

طرحهای بدون گروه کنترل

طرح تنها پس-آزمون تک-گروهی

طرح پیش-آزمون پس-آزمون تک گروهی

طرح مداخله حذف شده

طرح مداخله مکرر

طرحهایی که یک گروه کنترل دارند اما پیش-آزمون ندارند

طرح تنها پس-آزمون با گروههای غیرهم ارز

ارتقاء طرحهای بدون گروه کنترل از طریق مقابله دادن متضادهایی غیر از متضادهای همراه با گروه کنترل مستقل

طرح طولی یا پایش مورد

نتیجه گیری

۵. طرح های شبه-آزمایشی که هم گروه کنترل و هم پیش-آزمون را بکار می گیرند

طرحهایی گروه کنترل و پیش-آزمون دارند

گروه کنترل دستکاری نشده با نمونه های پیش-آزمون و پس-آزمون مستقل

جفت سازی در طول کنترل همگروهی ها

طرحهایی که عناصر متعدد طراحی را ترکیب می کنند

گروههای کنترل دستکاری نشده با پیش-آزمونها و پس-آزمونهای متعدد، متغییرهای غیرهم ارز وابسته، و مداخله های حذف شده و مکرر

ترکیب تکرارهای تبدیل شونده با یک طرح گروه کنترل غیرهم ارز

یک گروه کنترل دستکاری نشده به همراه پیش-آزمون دوبل و نمونه های مستقل و وابسته

عناصر طراحی

تخصیص

اندازه گیری

گروههای مقایسه

مداخله یا دستکاری

عناصر طراحی و انجام شبه-آزمایش ایده آل

نتیجه گیری

پیوست ۵.۱: پیشرفتهای مهم در زمینه تحلیل داده از طرحهای با گروههای غیر هم ارز

نمرات تمایل و سوگیری های پنهان

مدلسازی سوگیری انتخاب

مدلسازی معدلات ساختاری متغیر مکنون

۶. شبه-آزمایشها: طرحهای سری زمانی متوقف شده

سری زمانی چیست

تشریح انواع اثرات

پیشنهاداتی مختصر در باب تحلیل

سری زمانی متوقف شده ساده

تغییر در مقدار ثابت

تغییر در شیب

اثرات ضعیف و معوّق

تهدیدات رایج مرترتب بر روایی

اضافه کردن مولفه های دیگر به سری زمانی متوقف شده ساده

اضافه کردن یک سری زمانی با گروه کنترل دستکاری نشده

اضافه کردن متغیر مستقل غیرهم ارز

حذف مداخله در یک زمان مشخص

اضافه کردن تکرارهای متعدد

اضافه کردن تکرارهای تبدیل شونده

برخی مشکلات رایج همراه با طرحهای سری زمانی متوقف شده

مداخله های بطئی به جای مداخله های ناگهانی و یکباره

علّیت معوق

سری زمانی کوتاه

محدودیتهای همراه با بسیاری از داده های آرشیوی

پینهاداتی در باب سری های زمانی همزمان

نتیجه گیری

۷. طرحهای ناپیوستگی رگرسیونی

مبانی ناپیوستگی رگرسیونی

ساختار پایه

مثالهایی در مورد طرحهای ناپیوستگی رگرسیونی

ملزومات ساختاری طرح

تنوع ها در طرح پایه

نظریه طرح ناپیوستگی رگرسیونی

ناپیوستگی رگرسیونی به عنوان اثر مداخله در آزمایشهای تصادفی

ناپیوستگی رگرسیونی به مثابه مدلی کامل از فرایند انتخاب

پایبندی به نقطه برش

عدول یا ابطال نقطه برش

ریزش و جابجایی از یک مداخله به مداخله دیگر

ناپیوستگی رگرسیونی فازی

تهدیدات مترتب بر روایی

ناپیوستگی رگرسیونی و سری زمانی متوقف شده

روایی نتایج آماری و تعیین نادرست خصوصیات فرم کارکردی

روایی درونی

ترکیب آزمایشهای تصادفی با ناپیوستگی رگرسیونی

ترکیب ناپیوستگی رگرسیونی و شبه-آزمایش

ناپیوستگی آماری – آزمایش یا شبه-آزمایش

پیوست ۷.۱: منطق اثبات آماری در مورد ناپیوستگی رگرسیونی

۸. آزمایشهای تصادفی: منطق،طرحها، و شرایط لازم برای انجام آنها

نظریه تخصیص تصادفی

تخصیص تصادفی چیست؟

چرا تصادفی سازی اثرگذار و مفید است؟

تخصیص تصادفی و واحدهای تصادفی سازی

محدودیت در دستیابی به تخصیص تصادفی

برخی طرحهای قابل بکارگیری برای طرح های آزمایشی

طرح پایه

طرح گروه کنترل پیش-آزمون پس-آزمون

طرح مداخله جایگزین با پیش-آزمون

مداخلههای متعدد و کنترلهای با پیش-آزمون

طرحهای فاکتوریل

طرحهای طولی

طرح های جابجایی مداخله

شرایط مناسب برای تخصیص تصادفی

هنگامی که تقاضا بر عرضه پیشی می گیرد

هنگامی که یک نوآوری قابل ارایه به تمامی واحدها در آن واحد نیست

هنگامی که واحدهای آزمایشی می توانند به صورت موقت جداسازی شوند: طرح نمونه های-معادل-زمانی

هنگامی که واحدهای آزمایشی از نظر مکانی و فضایی مجزا هستند و مراودات و ارتباطات داخل واحد کم است

هنگامی که تغییر اجباریست و راه حلها شناخته شده فرض می شوند

هنگامی که یک رابطه می تواند شکسته شود یا ابهام در مورد نیاز می تواند حل شود

هنگامی که فرد هیچ گونه ترجیحی در مورد گزینه های مختلف ندارد

هنگامی که می توانید سازماندهی مختص خودتان را خلق کنید

هنگامی که بر واحدهای آزمایشی کنترل دارید

هنگامی که انتظار قرعه کشی می رود

زمانی که تخصیص تصادفی مطلوب یا ممکن نیست

بحث و نتیجه گیری

۹. استنباطهای تعمیم یافته علّی: روشهایی برای مطالعات متعدد

تعمیم بر اساس مطالعات منفرد در مقابل تعمیم بر اساس مطالعات متعدد

برنامه های مطالعات تحقیقی متعدد

مدلهای فازبندی شده برای مطالعات به طور افزایشی تعمیم پذیر

برنامه های هدایت شده آزمایشها

مرور بر متون تحقیقات موجود

مرور بر متون آزمایشها

مرور های متون تلفیق کننده تحقیقات آزمایشی و غیر آزمایشی

مشکلات همراه با مرور متون

مرور کمّی تحقیقات موجود

مبانی متا آنالیز

متا آنالیز و پنج اصل استنباط علّی تعمیم یافته

بحث در باب متاآنالیز

پیش‌گفتار

این کتاب مناسب آنهایی است که به این نتیجه رسیده اند که تشخیص یک رابطه وابستگی میان یک علّت و اثراتش یک اولویت است، همچنین کسانی که روشهای آزمایشی را برای بررسی این رابطه در نظر گرفته اند. اینگونه روابط علّی، در امور انسانی، اهمیت بسیاری دارند. پاداش شناسایی درست روابط علّی، و هزینه عدم شناسایی آنها هر دو می‌تواند بسیار زیاد باشد. دانستن اینکه آیا افزایش مدت تحصیل، سبب زندگی شادتر یا درآمد بیشتر در آینده خواهد شد، برای افرادی که در مرحله تصمیم گیری درباره صرف زمان بیشتر برای تحصیل در دانشگاه هستند، بسیار کاربردی خواهد بود. دانستن این امر همچنین به سیاست‌گزاران در تعیین میزان کمک مالی که باید به موسسات آموزشی مختلف تخصیص داده شود کمک خواهد کرد. از سال‌های اولیه وجود انسان، علّیت به شناسایی استراتژی‌هایی که در درمان بیماری موثر هستند، و داروهای موثر برای درمان، کمک کرده است. امروزه سودمندی روابط علّی به طور کامل درک شده است، به گونه‌ای که تلاش زیادی در حال انجام است تا این یافته‌ها هم در حالت کلی در امور انسانی، و هم به طور اخص در علم جایگاه خود را بیابند.

با این وجود، دانش گذشته به ما می آموزد که به ندرت پیش می‌آید علّتهایی در سراسر جهان، تحت همه شرایط و برای هر نوع انسان و در هر دوره تاریخی به صورت یکسان برقرار باشند. بدون شک همه ادعاهای علّی مشروط هستند. بنابراین اگرچه معمولاً تهدیدی خارج از گروه، سبب انسجام درون‌گروهی می‌شود، اما همواره این امر صحیح نیست. به عنوان مثال، در سال ۱۴۹۲، پادشاه گرانادا مجبور شد شاهد این باشد که افراد سپاه مغربیش حاضر به جنگ با سپاه پرتعداد پادشاهی کاتولیک اسپانیا که در نزدیکی سانتافی‌دلا فرونترا اتراق کرده بودند نشده، و شهر را ترک کرده و به خانه‌های اجدادی خود در شمال آفریقا رفتند. در اینجا، تهدید خارجی از سوی گروه اسپانیولی‌های مسیحی، سبب همبستگی اجتماعی بیشتر در اسپانیولی‌های مسلمان نشده بود، بلکه سبب جدایی نیروی دفاعی آنها شد. با این وجود برخی فرضیات علّی اقتضایی تر از دیگر فرضیات هستند، و مسلماً شناخت حداکثری این اقتضائات و روابطی که به طور باثبات‌تری برقرار است، بسیار سودمند است. به عنوان مثال، آسپیرین داروی بسیار قدرتمندی است زیرا نشانگان مربوط به بسیاری از بیماری‌های مختلف از جمله زکام، سرطان روده، و بیماری‌های قلبی عروقی را کاهش می‌دهد؛ این دارو چه در ارتفاع کم یا زیاد، چه در اقلیم‌های گرم یا سرد، به صورت کپسول یا مایع، برای کودکان و بزرگسالان عملکرد خوبی دارد؛ و برای کسانی که دچار مشکلات ثانویه‌ای غیر از زخم معده هستند، بسیار موثر است. اگرچه، طیف کاربرد داروهای دیگر بسیار محدودتر است. مثلاً تنها می‌توانند یک نوع سرطان را مداخله کنند، یا تنها در بیمارانی با درجه

خاصی از مقاومت فیزیکی کار می‌کنند، یا تنها زمانی که دوز کاملاً به درستی مصرف شود، عمل می‌کنند، یا تنها اگر پادتن‌هایی از پیش برای مقابله با بیماری در بدن تولید نشده باشد، عملکرد درستی دارند. اگرچه در بیان عامیانه اغلب به طور کلی از روابط علّی یاد می شود، اما شناسایی شرایطی که کاربرد این روابط را محدود می‌کند، همین فراوانی دارد.

این کتاب دو هدف اصلی دارد: شناسایی روابط علّی و درک تعمیم پذیری آنها. بنا به هدف اول راه‌هایی را پیشنهاد می‌کنیم که به کمک آنها می توان آزمون گزاره‌های علّی آزمایشی را در پروژه‌های تحقیقاتی بهبود بخشید. برای حصول این هدف، بهتر است به جای استفاده از رویه های مدلسازی آماری، از عناصر طراحی ساختاری قابل احصاء از نظریه آزمایش بهره بگیریم. پیشرفتهای آماری اخیر در زمینه‌ی استنباط علّی در داده‌های مشاهده‌ای (مانند Holland, 1986؛ Rosenbaum, 1995a؛ Rubin, 1986) سبب ارتقاء درک ما از این استنباطها شده است. با این وجود، اگر بخواهیم بر اساس تجربه خودمان از مشاوره‌هایی که به آزمایش‌های میدانی ارائه کرده‌ایم قضاوت کنیم، باید بگوییم که این پیشرفت‌ها بعضاً باعث ایجاد این انتظار غیرواقعی در میان برخی خوانندگان شده که کاربرد آمار جدید از جمله انطباق یا طبقه‌بندی نمرات تمایل، مدل‌های سوگیری انتخاب، و تحلیل‌های حساسیت سوگیریهای پنهان می‌تواند فی‌النفسه برای تضمین اعتبار استنباط علّی کافی باشند. اگرچه گاه این تعدیلها الزامی بوده و معمولاً بعد از اینکه عناصر طرح آزمایشی به خوبی طراحی و بکار گرفته شدند، مفید هستند، اما در غیاب عناصر طرح آزمایشی غالباً عملکرد ضعیفی دارند. خوشبختانه با گذشت زمان تاکید بیشتری بر ضرورت وجود این عناصر طراحی در آزمایشهای اقتصادی و علوم اجتماعی طراحی صورت می گیرد. مثلاً هکمان، ایشیمورا و تاد (Heckman, Ichimura and Todd, 1977) بر کاربرد چارچوب‌های اندازه‌گیری مشترک[1] و کنترل‌های محلی، وینشیپ و مورگان (Winship and Morgan, 1999) بر سودمندی وجود انواع پیش‌آزمون و پس-آزمونها، و روزنبام (Rosenbaum 1999b) بر اهمیت انتخاب‌های متعدد طراحی[2] در داده‌های مشاهداتی تاکید داشته‌اند. در تکمیل کار این محققین کتاب حاضر بر این نکته تاکید می کند که در قیاس میان طراحی و روشهای آماری، عناصر طراحی ارجحیت دارند (Shadish & Cook, 1999)!

دومین هدف کتاب حاضر، ارائه‌ی راههایی برای بهبود و ارتقاء تعمیم پذیری گزاره‌های علّی است. اگرچه رویه‌های نمونه‌گیری مرسوم، بهترین ابزارها برای تضمین تعمیم پذیری هستند، اما به ندرت می‌توان از آنها در تعمیم روابط علّی بهره گرفت. در عوض، ما در این کتاب به دنبال بهبود تعمیم علّی از طریق یک نظریه بنیادی تعمیم علّی هستیم. این نظریه، انعکاس دهنده اصولی است که آماردانان در جریان کارهای روزانه خود از آن بهره می‌گیرند تا بتوانند تعمیم‌هایی در حوزه‌های

[1] Common measurement framework
[2] Many design choices

متنوعی از جمله مدلسازی حیوانی بیماریهای انسانی، تصمیم‌گیری در باب اینکه آیا یک مورد به طبقه ای عمومی تر تعلق دارد، شناسایی روندهای کلی در مرور بر ادبیات، و تصمیم درباره اینکه آیا مطالعات همه‌گیرشناسی رابطه کلی بین دود سیگار و سرطان را پشتیبانی می‌کنند یا نه، انجام دهند. امیدواریم که نتیجه نظریه ای در باب تعمیم روابط علّی باشد که از نظریه نمونه گیری کاربردی تر باشد، ولی در عین حال نظریه نمونه گیری را به عنوان یک مورد خاص در بر بگیرد.

این کتاب در ادامه‌ی کتاب‌های کمپبل و استنلی (Campbell and Stanley, 1963) و کوک و کمپبل (Cook and Campbell, 1979)، و برای پیگیری دو هدف فوق الذکر منتشر می‌شود. با این وجود، کتاب حاضر از چند جهت با دو کتاب یاد شده تفاوت دارد. مشهودترین تفاوت این است که در اینجا ما بر تعمیم روابط علّی تاکید داریم. اگرچه در کارهای گذشته نیز بر اهمیت این تعمیم تاکید شده است، و حتی اصطلاح «روایی بیرونی» نیز برای بیان آن ابداع شده است، اما تاکید بیشتر بر آن بود که بررسی شود آیا یک رابطه خاص مورد آزمون در یک زمینه تحقیقاتی خاص علّی است یا نه. در این کتاب، روش‌های مطالعه روایی بیرونی مورد توجه ویژه قرار می‌گیرند، در صورتی که در کتابهای پیشین، روش‌های توجه بیشتر به روایی درونی معطوف شده بود.

اختلاف دوم در این است که ما ناچار به دست و پنجه نرم کردن با فلسفه علم متاخر بودیم. فلسفه ای که برخی از اصلی ترین ستونهایی که ساختار منطق علمی بر آنها استوار است را زیر سوال می برد- به ویژه آنجا که به احتمال عینیت (بی غرضی) و جایزالخطا بودن قیاسها و استقرا به عنوان راههایی برای کسب دانشی خاص مربوط می شود. همچنین، پیامدهای بسیاری از یافته‌های توصیفی فراعلم (مطالعه سیستماتیک تاریخ، جامعه‌شناسی، روان‌درمانی، و فلسفه علم) از این دسته هستند، که در موارد فراوانی، اقدامات علمی آنها با منطق علمی مرجح امروزی تفاوت دارد. علم توسط انسان‌ها ایجاد می‌شود و توسط گروهی از دانشمندان که علایق شناختی و اقتصادی به تعریف، دفاع و ارتقای علم دارند، اعتبارسنجی می‌شود. بنابراین، این کتاب حتی بیشتر از کتاب‌های پیشین، امکان خطا را مفروض می داند. اما هرگز به خاطر این باور، آن را کنار نمی‌گذاریم و از سوی دیگر هم نمی‌گوییم که همه چیز درست است. ماهیت خطاپذیر علم منجر به بی‌ارزش بودن (یعنی اگر کامل نیست، پس ارزشی ندارد) یا نسبیت‌گرایی قوی در روش‌شناسی (یعنی هیچ روشی هرگز نسبت به روش دیگر، با هیچ هدفی برتری نداشته است) نمی‌شود. بلکه ما از این باور دفاع می‌کنیم که برخی از گزاره های علّی بهتر از دیگر گزاره ها قابلیت تضمین دارند و منطق و تجربه عملی در علم بیانگر این است که برخی از اقدامات، از نظر اهداف علّی اغلب (نه همیشه) برتر از دیگر اقدامات هستند؛ اگرچه لزوماً برای اهداف دیگر اینگونه نیست.

اختلاف سوم این است که به جای تاکید بر طراحی، بر عناصر طراحی تاکید می‌شود؛ به ویژه زمانی که مطالعات تجربی بدون تخصیص تصادفی به شرایط مداخله انجام می‌شوند. اقدام علمی‌ای که بیشترین همبستگی را با تحقیقات علّی دارد، آزمایش است، که تمامی اشکال آن به عنوان هدف اصلی این کتاب در نظر گرفته می‌شود. امروزه، منظور از آزمایش، مطالعه سیستماتیک طراحی شده برای بررسی نتایج تغییر عامدانه یک عامل علّی بالقوه است. متخصصان آزمایشگاهی، نیازمند اقدامات زیر هستند: (۱) تغییر دادن مداخله، (۲) اندازه‌گیری نتایج پس از مداخله، (۳) حداقل یک واحد که مشاهده روی آن انجام شود و (۴) مکانیسمی برای استنباط اینکه نتایج بدون مداخله چه خواهد بود (استنباط خلاف واقع نامیده می‌شود، یعنی ما استنباط می‌کنیم که مداخله اثری تولید می‌کند که در صورت عدم وجود مداخله، آن اثر وجود نمی داشت). خواهیم دید که ویژگی‌های ساختاری دیگر زیادی برای آزمایش وجود دارد که اغلب آنها برای بهبود کیفیت استنباط خلاف واقع به کار گرفته می‌شوند. اگرچه علیرغم محبوبیتی که آزمایش‌ها در علوم طبیعی، علوم ریاضی، پزشکی، روانشناسی، آموزش و اقتصاد کار دارند، تنها روش تحقیقاتی که می‌تواند نتایج علّی را اثبات کند، نیستند. بسیاری از مطالعات همبستگی در جامعه‌شناسی، علوم سیاسی، علوم توسعه‌ای و شاخه‌های خاص اقتصاد بر ایده‌های علّی متکی هستند، که برای توسعه نظریه ارائه شده‌اند، اما از ساختار یا زبان رسمی آزمایش بهره نمی‌گیرند. همه روش‌های غیرآزمایشی را می‌توان از نظر مولفه‌های طراحی ساختاری که در آنها وجود دارد یا وجود ندارد بررسی کرد، و ضعف‌ها و قوت‌های آنها در استنباط علّت را مشخص نمود. ما بر این باوریم که بهتر است به جای در نظر گرفتن یک سری معین از طرح‌ها، با توصیف مولفه‌های ساختاری که مشخصه آزمایشها هستند و نشان دادن نحوه ترکیب آنها برای خلق طرح‌های آزمایشی که تا پیش از این مورد استفاده قرار نگرفته‌اند، مولفه‌های ساختاری آزایشها را در نظر بگیریم. این طرح‌ها اساس کتاب‌های پیشین هستند (Campbell & Stanley, 1963؛ Cook & Campbell, 1979). در مقابل، ما با تمرکز بر مولفه‌های طراحی به دنبال این هستیم که به خوانندگان کمک کنیم تا مجموعه ابزارهایی بدست آورند که آنقدر منعطف باشد که برخی از آنها را بتوان برای بهبود گزاره های علّی در هر زمینه تحقیقاتی مورد استفاده قرار داد.

اختلاف چهارم بین این کتاب و کتاب کوک و کمپبل (Cook and Campbell, 1979) در این است که در اینجا چندان به تحلیل آماری داده‌ها پرداخته نمی‌شود. به جای آنکه فصول کاملی راجع به جزئیات آماری ارائه شوند، پاراگراف‌هایی کوتاه یا بخش‌هایی را به تحلیل داده ها اختصاص داده و بیشتر توجه خود را معطوف به مفاهیم کرده ایم. اگرچه به طور کم شمار به معادلاتی در متن یا در پاورقی اشاره می کنیم، و خواننده را برای مطالعه بیشتر به منابع دقیق‌تر ارجاع می‌دهیم. دلیل این کار تا حدی کاربردیست. بیست سال پیش، توصیفات قابل فهم رویه های آماری مانند سری

زمانی یا طرح‌های گروه کنترل غیرهم‌ارز آنقدر کمیاب بود که راه حلها (درمانهای) بسط یافته مورد نیاز بود. اما امروزه، راه حل های آماری این موضوعات به طور گسترده برای هر سطح فنی ایجاد شده است. بنابراین بهتر دانستیم فضای موجود را به مفاهیم مربوط به طراحی و تعمیم اختصاص دهیم. اگرچه، کاهش توجه به آمار، بیانگر ترجیح راه حل‌های طراحی نسبت به راه حل‌های آماری برای استنباط علّی نیز هست (به همه دلایل فوق‌الذکر).

اختلاف پنجم در این است که این کتاب، اصلاحاتی بر طرح مفهومی کلی‌ای که همواره در طول سالها مرکز توجه کارهای کمپل بوده است، یعنی گونه شناسی روایی صورت داده است. تغییرات در اغلب موارد حداقلی هستند و ما توجه خود را همچنان بر چهار نوع روایی (درونی، نتایج آماری، سازه و بیرونی) و بر مرکزیت شناسایی تهدیدهای موجه روایی در استنباطهای علّی کاربردی معطوف می داریم. اما چارچوب روایی را به چند طریق تغییر داده‌ایم. مثلاً در روایی نتایج آماری، سعی کرده‌ایم که به جای توجه بیش از اندازه به معناداری آماری، وزن بیشتری برای بزرگی (اندازه) یک اثر قائل شویم. تفکر ما در باب تعمیم (روایی درونی و سازه)، منعکس‌کننده اثر مطالعات منسجم کرنباخ (Cronbach, 1982) بر روی مسائل تعمیم علّی است. همچنین تغییرات کوچکی هم در فهرست تهدیدهای روایی انجام داده‌ایم. اگرچه بسیاری از این تغییرات تنها مورد علاقه نظریه‌پردازان روش شناسی آزمایشی است، اما همچنان امیدواریم که برخی از آنها (مانند افزایش تاکید بر اندازه اثر) اثری کاربردی بر آزمایشها در عمل داشته باشد.

علیرغم تغییرات ذکر شده، تاکید کلی این کتاب، بر آزمایشهای میدانی، و بر رفتار انسان در مختصات غیرآزمایشگاهی است (اگرچه بیشتر این کتاب به آزمایش‌های آزمایشگاهی می‌پردازد). در مکان‌هایی مانند مدارس، دفاتر تجاری، کلینیک‌ها، بیمارستان‌ها، سازمان‌های رفاهی، و خانه‌ها، محقق فاصله بسیار زیادی از دستیابی به کنترل کامل (بر عناصر آزمایش) دارد، معمولاً مهمان است، و باید مذاکره کند و نه اینکه دستور بدهد، و معمولاً به جای اینکه هر چیزی را که می‌خواهد بدست بیاورد، باید مصالحه کند. انجام برخی از مصالحه‌ها، نگران‌کننده‌تر است. به ویژه آزمایشگران میدانی علاقه‌ای ندارند که از کنترل روی اندازه‌گیری، انتخاب و فرایند زمانبندی مداخله به ویژه در انتساب مداخله صرفنظر کنند. زیرا زمانی که افراد به صورت کامل به انتخاب خودشان به شرایط مداخلات مختلف تخصیص داده می شوند، استنباط علّی بسیار دشوار می شود. با این وجود، بدیهی است که این کنترل، معمولاً به جای اینکه تصمیمی یک طرفه باشد، موضوعی برای مذاکره و مصالحه است.

در نهایت می‌خواهیم از همراهی نویسنده سوم این کتاب، دونالد توماس کمپبل تشکر کنیم که در می ۱۹۹۶ و زمانی که تنها نیمی از این کتاب آماده شده بود، درگذشت. سپاسگزاری از نقش او

کار آسانی نیست. مسلماً نقش او فراتر آنچه که برای این کتاب نوشت، بوده است. ایده‌های او به صورت گسترده و عمیق بر همکاران و دانشجویانش تاثیر گذاشته است. او بنیان‌گزار تمامی سنت آزمایش میدانی و شبه آزمایش ارائه شده در این کتاب است و این سنت آنچنان به او وابسته بود که ما و همکاران دیگرمان، گاهی آن را کمپبلی می‌نامیدیم. بسیاری از مهمترین مفاهیم این کتاب مانند روایی درونی و بیرونی، تهدیدهای روایی و منطق آنها، و شبه آزمایش، توسط او ابداع شده و بسط یافتند. بسیاری از ایده‌های دیگر او، درباره خطاپذیربودن شکلگیری دانش (ما خویشاوندان آمیب‌ها هستیم و هیچ الهام مستقیمی که در آن با او مشترک نباشیم دریافت نکرده ایم. چگونه می‌توانیم تا این حد مطمئن باشیم؟)، درباره ماهیت متغیر و تصادفی پیشرفت علمی و درباره ماهیت اجتماعی سرمایه‌گذاری علمی، بخش‌هایی از تفکر ما هستند که به صورت ضمنی در سراسر این کتاب دیده می‌شوند. دین ما به کمپبل به عنوان همکار و به عنوان دانشجو، بدون شک فراتر از آن است که بتوانیم بیان کنیم.

کمپبل (۱۹۸۸) عاشق یکی از استعاره هایی بود که اغلب توسط دبلیو. وی. کوئین (فیلسوف و ریاضی‌دان) مورد استفاده قرار می گرفت. او می‌گفت که دانشمندان مانند ملوانانی هستند که باید یک کشتی فرسوده را در دریا تعمیر کنند. آنها وقتی مجبورند یک تخته ضعیف را جایگزین کنند، به حجم بزرگی از الوار اعتماد می‌کنند. هر کدام از این الوارها که اکنون مورد اعتماد است، در زمان خود با الوار دیگری جایگزین خواهند شد. نسبت تخته هایی که جایگزین می شوند به آنهایی که محکم تلقی می شوند باید همواره کوچک باشد. کمپبل از این استعاره برای نشان دادن نقش فراگیر اعتماد در علم، و فقدان بنیانهای به واقع مستحکم در علم بهره می گرفت. چهار مصرع زیر که مربوط به شعر «بستر مستقر[۳]» سیموز هینی (۱۹۹۱) هستند، نه تنها خلاصه‌ای از عشق کمپبل به استعاره کوئین را نشان می‌دهند بلکه سهم خود کمپبل را در یکی از کشتی‌های علم بیان می‌کنند:

و اکنون، این میراث است
شریف، بدوی، با تخته‌پوشی صلب
آمده از گذشته ای دور، اما با اشتیاق رو به جلو
باز هم و باز هم و باز هم.[۴]

[3] Settle bed

[۴] برگرفته از «بستر اقامت» از کتاب زمین باز: شعرهای انتخابی ۱۹۶۶ ۱۹۹۸ توسط سیموز هینی.

همچنین برای خوانندگانی که زبان عامیانه‌تر را دوست دارند، ما از عبارات وودی گوتری[5] بهره می‌گیریم که در آهنگ «مرد دیگری بر باد رفت[6]» به عنوان پیش‌بینی مرگش نوشته است: «نمی دانم، ممکن است به بالا یا پایین یا هر جای دیگری بروم، اما احساس می‌کنم ممکن است این خط خطیها بر جای بماند». امیدواریم این کتاب کمک کند تا نقش ارزنده دان در عرصه ی مطالعات آزمایشی میدانی برای نسل‌های آینده، حفظ شود.

ویلیام آر. شدیش
توماس، دی کوک

[5] Woody Guthrie
[6] Another Man’s Done Gone

فصل ۱

آزمایش‌ها و استنباط علّی تعمیم یافته[۷]

آزمایش (Ik-sper Ə-mənt) (کلمه از ریشه زبان انگلیسی میانه (بین حدود ۱۱۰۰ تا ۱۵۰۰ میلادی)، از فرانسه قدیم و لاتین کلمه experimentum، که از ریشه کلمه experiri به معنای امتحان کردن، گرفته شده است): ۱. الف. آزمونی که تحت شرایط کنترل شده برای به تصویر کشیدن حقیقتی معلوم ، بررسی روایی و درستی یک فرضیه و یا تعیین کارآمدی چیزی که تا پیش از این مورد آزمون قرار نگرفته، طراحی می‌شود. ب. فرایند انجام چنین آزمونی (experimentation) ۲. رویه یا عملی نوآورانه

علّت: (Cause (koz (کلمه از ریشه زبان انگلیسی میانه، از فرانسه قدیم و از ریشه لاتین کلمه causa، به معنی دلیل و هدف گرفته شده است): ۱. تولید کننده یک اثر، نتیجه یا عاقبت. ب. عاملی مانند یک فرد، رویداد، و یا شرایط که مسئول بوجود آمدن یک نتیجه یا فعالیت است. ۲. سبب شده یا تحریک کردن از طریق اجبار یا نفوذ

در نظر بسیاری از فلاسفه و مورخین، تأکید فزاینده بر آزمایش در قرنهای ۱۶ و ۱۷ نشان‌دهنده ظهور علم مدرنی بود که ریشه در فلسفه طبیعی داشت (Hacking, 1983). درک (Drake, 1981) رساله ۱۶۱۲ گالیله در باب اشیاء شناور روی سطح و غوطه‌ور در آب را، به عنوان طلیعه علم آزمایش مدرن معرفی می‌کند. اما مطالعه ویلیام گیلبرت (William Gilbert, 1600) بر روی گرانش و اشیاء آهنربایی، آزمایشات متعدد لئوناردو داوینچی، و شاید حتی تلاش‌های امپیدوکلس فیلسوف قرن پنجم پیش از میلاد را می‌توان همچنان به عنوان نمونه‌های قدیمی‌تر علم آزمایشی بر شمرد (Jones, 1969a, 1969b). در باب معنای روزمره کلمه آزمایش نیز می‌توان گفت بشر از

7 Experiments and Generalized Causal Inference

اولین حرکاتش در تاریخ، راه‌های مختلف انجام امور را به طور مستمر می‌آزماید. این نوع از آزمایش کردن بخشی طبیعی از زندگی ماست، مانند امتحان کردن دستورهای پخت جدید برای یک غذا و یا پیدا کردن راهی متفاوت برای روش کردن آتش در یک پیک‌نیک.

با این وجود، انقلاب علمی‌قرن هفدهم از سه جهت راهش را از کاربرد عمومی‌مشاهدات در فلسفه طبیعی مورد استفاده در آن دوران جدا کرد. اول، مشاهده را به طور فزاینده ای برای اصلاح خطاهای موجود در نظریه‌ها بکار گرفت. در طول تاریخ، فیلسوفان طبیعت‌گرا، برای پیروزی در بحث‌های فلسفی معمولاً با یافتن مشاهداتی که از نظریه آنان پشتیبانی می‌کرد، از مشاهده در نظریه‌های خود استفاده می‌کردند. با این حال، آنان هنوز هم به استفاده از مشاهده برای استنباط نظریه‌ها از "اصول اولیه" -نقاط شروعی که انسان‌ها بنابر ماهیت ما یا وحی الهی، می‌دانند واقعیت دارند (مثلاً ، ویژگی های مفروض چهار عنصر اصلی آتش، آب، خاک و هوا در فلسفه طبیعی ارسطویی)- وابسته اند. بر اساس برخی روایات، این وابستگی نظریه به شواهد تجربی، در قرن هفدهم رو به انحطاط گذاشت: اصل ارسطویی برتری تجربه[8] در میان فلاسفه رو به زوال گذاشت و جای خود را به تکیه کردن بر استدلال‌های مبتنی بر مثال‌های علَی و رد نقیض‌ها -با اشاره به استثنائات آشکار و واضحی که هنوز مورد آزمون قرار نگرفته‌اند- داد (Drake, 1981, p. xxi). زمانی که برخی از متفکرین قرن هفدهمی‌شروع به استفاده از مشاهده برای اصلاح خطاهای مسلم و آشکار در اصول و مبانی اولیه نظری و مذهبی، کردند؛ این تلاش منجر به تعارض میان آنها و مراجع دینی و فیلسوفان شد (همانطور که در مورد اجبار گالیله برای پس گرفتن نظریه اش در خصوص گردش زمین به دور خورشید رخ داد). با در نظر گرفتن این مخاطرات، این حقیقت که علوم تجربی جدید تعادل را به سمت توجه به مشاهده و اجتناب از جزم‌اندیشی متمایل کرده، قابل توجه است. تا زمان مرگ گالیله، نقش مشاهدات نظام‌مند به عنوان عنصر اصلی علم تثبیت شده، و از آن زمان تا به حال این نقش پایدار مانده است.

دوم، قبل از قرن هفدهم گرایش به تجربه عموماً بر پایه مشاهده غیرفعال سیستم‌های درحال فعالیت بود، و نه بر پایه مشاهده آنچه پس از تغییر حساب‌شده سیستم رخ خواهد داد. پس از انقلاب علمی‌در قرن هفدهم، کلمه آزمایش به معنای انجام اقدامی‌حساب شده، که متعاقب آن رشته‌ای سیستماتیک از مشاهدات در مورد آنچه رخ خواهد داد، انجام خواهد شد، به کار گرفته شد. همانطور که هکینگ (Hacking, 1983) در مورد فرانسیس بیکن می‌گوید: او بر این باور است که «هرچند باید طبیعت را به طور بکر و دست نخورده مشاهده کنیم، اما باید دم شیر را نیز ببینیم؛ به این معنی که باید دنیای خود را دستکاری کنیم تا رازهای آن را دریابیم (ص. ۱۴۹).»
اگرچه مشاهده غیرفعال حقایق زیادی را در مورد جهان پیرامون برای ما روشن می‌سازد، اما دستکاری فعال لازم

8 Appealing to experience

است تا بتوان برخی قوانین و ممکنات را کشف کرد. به عنوان مثالی پیش‌پا افتاده، فولاد ضدزنگ[9] را در نظر بگیرید. فولاد ضد زنگ به طور طبیعی به وجود نمی‌آید، بلکه انسان باید دستکاری‌هایی انجام دهد تا آن را خلق کند. مشاهده اثر چنین دستکاری‌هایی موضوع توجه علوم تجربیست.

سوم، آزمایش‌کنندگان از همان ابتدا مطلوبیت و اهمیت کنترل اثرات عوامل بیرونی که ممکن است بتوانند مشاهدات را محدود کرده، و یا نتایج را به انحراف کشانده و دچار سوگیری نمایند، را دریافتند. بنابراین تلسکوپ‌ها به بالاترین نقاط انتقال داده شدند، مکان‌هایی که در آنجا آسمان و هوا شفافتر بود، دقت لنز تلسکوپ‌ها هرچه بیشتر تقویت شد، و دانشمندان آزمایشگاه‌هائی را ساختند که در آنها امکان بکارگیری دیوارها برای بیرون نگه‌داشتن سوگیری‌های بالقوه موج‌های اتر وجود داشت و می‌توانستند در آن آزمایشگاه‌ها از لوله‌های آزمایش استرلیزه شده‌ای که از گزند آلودگی و باکتری در امان مانده بود استفاده کنند. در ابتدا این کنترل‌ها در زمینه‌هایی همچون ستاره‌شناسی، شیمی و فیزیک ابداع شد، یعنی همان علوم تجربی که اولین بارقه‌های علم در آنها دیده شده بود. اما هنگامی‌که محققین تلاش کردند تا آزمایش‌ها را در زمینه‌هایی همچون بهداشت عمومی یا آموزش بکارگیرند، دریافتند که کنترل اثرات خارجی در این زمینه‌ها بسیار دشوارتر است و کنترل‌های مورد استفاده در آزمایشگاه‌ها برای کاربردهای جدید ناکارآمد است. بنابراین روش‌های جدیدی برای مواجهه با اثرات عوامل بیرونی ابداع شد، که از آن جمله می‌توان به تخصیص تصادفی[10] و یا افزودن گروه‌های کنترل غیرتصادفی به آزمایش اشاره داشت. با افزایش روزافزون تجربیات مشاهده‌ای و نظری در باب این موضوعات، منابع سوگیری بیشتری شناسایی، و روش‌های بیشتری برای مقابله با آنها طراحی شد.

امروزه، عنصر مشترک تمامی‌آزمایش‌ها همچنان تغییر حساب‌شده چیزی، برای کشف اتفاقاتی است که برای چیزی دیگر رخ می‌دهد (کشف اثرات علّت مفروض). به عنوان یک فرد عادی نیز در زندگی روزمره این کار را انجام می‌دهیم. مثلاً هنگامی‌که اثرات ورزش کردن را روی فشار خون خودمان ارزیابی می‌کنیم، و یا اثر رژیم گرفتن کمتر را روی افزایش وزنمان مشاهده می‌کنیم. این فصل با بحث در خصوص این موضوعات آغاز می‌شود. در ابتدا (۱) ماهیت علّیتی که بواسطه آزمون تجربی بررسی می‌شود را مورد بحث قرار خواهیم داد، سپس (۲) اصطلاحات تخصصی (مانند آزمایش‌های تصادفی و شبه‌آزمایش‌ها) که تبیین‌کننده آزمایش‌ها در علوم اجتماعی هستند را توضیح خواهیم داد. پس از آن (۳) مسائل همراه با چگونگی تعمیم روابط علّی بدست‌آمده از یک آزمایش را معرفی می‌کنیم، و در نهایت (۴) به طور خلاصه جایگاه آزمایش را درون بدنه بزرگ ادبیات مرتبط با ماهیت علم بیان خواهیم کرد.

9 Stainless steal

10 Random assignment

آزمایش‌ها و علّیت

بحث محسوس و قابل فهم برای خوانندگان در باب آزمایش‌ها نیازمند وجود یک دایره لغات مشترک در مورد علّیت، و درک درست مفاهیمی‌است که مبنای این دایره لغات را تشکیل می‌دهند.

تعریف علّت، اثر و روابط علّی

اغلب افراد معنای روابط علّی را به طرز مشهودی در زندگی روزمره خود تشخیص می‌دهند. برای مثال، می‌گوییم که اتومبیلی که با ماشین شما تصادف کرده است، علّت آسیب وارده به ماشین شماست؛ تعداد ساعاتی که صرف مطالعه کرده‌اید علّت نمرات کسب‌شده در امتحان است؛ و اینکه مقدار غذایی که فردی مصرف می‌کند علّت وزن زیاد وی است. ممکن است فرد حتی تا استنباط روابط پیچیده‌تر علّی نیز پیش رفته و چنین استنباط شود که کسب نمرات پایین در آزمون موجب ناراحت‌شدن و تخریب روحیه فرد شده، که خود منجر به کاهش میزان مطالعه، و در نتیجه منجر به کسب نمرات پایینتر می‌شود. در اینجا یک متغیر (نمره پایین) می‌تواند هم علّت باشد، و هم اثر (هم معلول). همچنین می‌تواند یک رابطه دوطرفه[11] میان دو متغیر وجود داشته باشد (نمرات پایین و مطالعه نکردن)، به‌گونه‌ای که هر یک موجب بوجود آمدن دیگری می‌شود.

علیرغم وجود چنین آشنایی شهودی با روابط علّی، تعریف دقیق علّت و اثر برای قرن‌ها فلاسفه را سردرگم کرده است. یقیناً اصطلاحاتی مانند علّت و اثر تا حدودی وابسته به یکدیگر، و وابسته به روابط علّی‌ای است که محمل علّت و اثر هستند. بر این اساس، جان لاک فیلسوف قرن هفدهم ادعا می‌کند « آن چیزی که ایده‌ای ساده یا پیچیده را می‌سازد، علّت نامیده می‌شود. و آن چیزی که تولید می‌شود، اثر نامیده می‌شود (John Locke, 1975, p.324)». از آن زمان تا کنون، دیگر فلاسفه و دانشمندان تعاریف سودمندی در مورد علّت و اثر و رابطه علّی ارائه کرده‌اند. تعاریفی که دقیقتر بوده و چگونگی کار آزمایش را به شکل بهتری روشن می‌سازند.

علّت

علّت آتش‌سوزی در یک جنگل را در نظر بگیرید. می‌دانیم که آتش به شیوه متفاوتی شروع می‌شود (مثلاً یک کبریت یا ته‌سیگاری که از ماشین بیرون انداخته می‌شود یا آتش نیمه‌روشن گردشگران). هیچکدام از این دلایل برای آتش‌سوزی جنگل ضروری نیست، چون یک جنگل بدون حضور این عوامل نیز می‌تواند دچار حریق شود. همچنین هیچکدام از این دلایل برای روشن شدن آتش کافی نیستند. با تمام این احوال، یک کبریت باید به مدت کافی داغ بماند تا احتراق آغاز شود، این کبریت باید در تماس با مواد قابل احتراق مانند برگ‌های خشک قرار داشته باشد، باید اکسیژن وجود داشته باشد تا احتراق اتفاق بیافتد، و هوا باید به اندازه کافی خشک باشد، به نحوی که برگ‌ها خشک بوده و کبریت بوسیله نم و باران مرطوب نشود. بنابراین کبریت بخشی از مجموعه

11 reciprocal

شرایطی است که در غیاب آنها آتشی روشن نخواهد شد. اگرچه وجود برخی از شرایط و پیش‌فرض‌ها مانند اکسیژن را می‌توان مفروض دانست. بنابراین یک کبریت روشن را می‌توان بخشی ناکافی اما غیراضافی از مجموعه شرایطی کافی اما غیرضروری دانست، این همان چیزیست که مکی (Mackie, 1974) از آن به عنوان شرط ضروری[۱۲] یاد می‌کند. ناکافی است چون کبریت در غیاب دیگر عوامل احتراق نمی‌تواند باعث روشن شدن آتش شود. غیراضافی است، تنها اگر، بتواند چیزی را به مجموعه عوامل احتراق اضافه کند که قابل تأمین توسط دیگر عوامل احتراق‌آور مانند اکسیژن و برگ خشک (برای شروع آتش) نباشد. با تمام این احوال، اگر فردی همزمان سعی کرده باشد تا با یک فندک آتش را روشن کند، همچنان دشوار خواهد بود که بگوییم که کبریت علّت آتش بوده است. کبریت در ترکیب با مجموعه کاملی از عوامل، بخشی از شرایط کافی برای شعله‌ور کردن یک آتش را شکل می‌دهد. اما این شرط ضروری نیست چون دیگر مجموعه‌هایی وجود دارند که می‌توانند آتش را بوجود بیاورند.

به عنوان مثال پژوهشی از شرایط ضروری، به درمان جدیدی برای سرطان اشاره می‌کنیم. در اواخر دهه نود تیمی‌از محققین به سرپرستی دکتر فلکمن اعلام کردند که داروی جدیدی به نام اندوستاتین ساخته‌اند که از طریق محدود کردن رگهای خونرسان به غده‌های سرطانی، باعث کوچک شدن این تومورها می‌شود (Folkman, 1996). دیگر محققان سرشناس تلاش کردند تا با تکرار آزمایش موردنظر اثر را تولید کنند، اما موفق به انجام این کار نشدند- حتی زمانی که دارویی ارسال شده از آزمایشگاه دکتر فلکمن را مورد استفاده قرار دادند. اما هنگامی‌که همین دانشمندان به آزمایشگاه دکتر فلکمن سفر کرده، و آزمایش را در آزمایشگاه او تکرار کردند، توانستند نتایجی مشابه نتایج دکتر فلکمن بدست آوردند. یکی از حاضرین، شرایط موجود در آن آزمایشگاه را پدیده «در دستان ما[۱۳]» نامید. به این معنی که حتی خود ما نیز نمی‌دانیم کدامیک از جزئیات اهمیت دارند، بنابراین احتمالاً مدت زمانی باید وقت صرف کنیم تا نتیجه بگیریم و بتوانیم این کار را انجام دهیم. اندوستاتین یک شرط ضروری بود. این دارو فی‌النفسه ناکافی بود، و برای اثربخش بودن لازم بود تا در مجموعه‌ای از شرایط مستتر باشد. شرایطی که محققین اولیه نتوانسته بودند به درستی آن را درک کنند.

غالب علّتها در واقع شرط ضروری هستند. معمولاً عوامل زیادی برای ایجاد شدن یک اثر باید وجود داشته باشد، اما محققین به ندرت تمامی‌آنها و روابط متقابل میان آنها را می‌شناسند. این یکی از دلایلی است که باعث می‌شود روابط علّی مورد بحث در این کتاب را قطعی[۱۴] قلمداد نکرده، و تنها بتوان ادعا کرد افزایش‌دهنده احتمال وقوع یک اثر هستند (Eells, 1991; Holland, 1994). این موضوع همچنین روشن می‌کند که چرا یک رابطه علّی ممکن است تحت برخی شرایط اتفاق بیافتد اما بطور عمومی و در زمانها، مکان‌ها، جمعیت‌ها و یا

12 Inus condition
13 In-our-hands
14 deterministic

مداخله‌ها و متغیرهای نتیجه‌ای دیگری که تقریباً شبیه شرایط اولیه هستند، به وقوع نپیوندند. تمامی روابط علّی، با درجاتی متفاوت، زمینه-محور یا وابسته به زمینه[۱۵] هستند، و بنابراین تعیم اثرات آزمایشی همواره یک مسأله است. و در حقیقت این دلیل اصلی پرداختن به این نوع تعمیم‌ها در کتاب حاضر است.

اثر

برای درک بهتر چیستی اثر، از مدل خلاف واقع[۱۶] دیوید هیوم (Lewis, 1973) فیلسوف قرن هجدهم استفاده می‌کنیم. خلاف واقع چیزی است که در تناقض با اثر باشد (در خلاف جهت اثر باشد). در یک آزمایش آنچه مورد مشاهده قرار می‌گیرد این است که هنگامی‌که افراد درمان یا مداخله را دریافت می‌کنند چه اتفاقی می‌افتد؟ خلاف واقع حالتی را نشان می‌دهد که برای همان افراد می‌توانست رخ دهد، اگر درمان موردنظر را دریافت نمی‌کردند. اثر عبارت است از تفاوت میان آنچه رخ داده و آنچه می‌توانست رخ دهد.

البته در حقیقت نمی‌توان یک خلاف واقع را مشاهده کرد. بیماری فنیل کتونوری (PKU) را در نظر بگیرید. بیماری متابولیکی که موجب عقب‌ماندگی ذهنی در نوزادان می‌شود، مگر آنکه بتوان بیماری را در هفته‌های اولیه زندگی نوزاد تشخیص داد. بیماری PKU در واقع ناشی از نبود آنزیمی‌در بدن است که از تشکیل فنیل‌آلانین — عنصری سمّی برای دستگاه عصبی- در بدن جلوگیری می‌کند. اگر رژیم بدون فنیل‌آلانین از مراحل اولیه زندگی نوزاد آغاز شود، از عقب‌ماندگی ذهنی جلوگیری خواهد شد. در این مثال، علّت را می‌توان نقص ژنتیکی، اختلال آنزیمی و یا رژیم غذایی در نظر گرفت؛ و در نظر گرفتن هر کدام از آنها متناظر خلاف واقع متفاوتی خواهد بود. برای مثال، اگر بگویید که رژیم غذایی فاقد فنیل‌آلانین باعث کاهش عقب‌ماندگی ناشی از فنیل‌کتونوری می‌شود، بنابراین خلاف‌واقع عبارت از این خواهد بود که اگر همین نوزادان رژیم غذایی بدون فنیل‌آلانین دریافت نمی‌کردند، چه اتفاقی برایشان می‌افتاد. همین منطق در مورد اختلال آنزیمی و ژنتیکی نیز مصداق دارد، اما غیرممکن است که یک نوزاد در آن واحد هم رژیم داشته باشد، و هم نداشته باشد، هم دچار اختلال ژنتیکی و آنزیمی‌باشد، و هم نباشد.

بنابراین، یکی از وظایف اصلی تمامی پژوهش‌های علّت-یابی[۱۷]، ایجاد نظیرهایی[۱۸] برای این خلاف‌واقع‌های غیرممکن است. برای مثال، اگر از نظر اخلاقی صحیح بود، می‌شد به تعدادی از نوزادان فنیل‌کتونوری رژیم داده، و به تعدادی دیگر از نوزادان فنیل که از نظر خصوصیات (سن، جنسیت، نژاد، مشخصات، وضع سلامتی، خانوادگی و اجتماعی) مشابه گروه اول هستند، رژیم داده نشود، و سپس نتایج در دو گروه با یکدیگر مقایسه شود. یا اگر اخلاقی بود، می‌توانستیم نتایج نوزادانی که برای سه ماه اول زندگی خود تحت رژیم نبوده اند را با

15 Context-dependent
16 counterfactual
17 Cause-probing
18 Approximation

نتایج بدست‌آمده از همان نوزادان پس از آنکه در ماه چهارم تحت رژیم فنیل آلانین قرار گرفتند مقایسه کنیم. هیچکدام از این دو نظیر یک خلاف‌واقع حقیقی نیستند. در مورد اول، هر کدام از نوزادان در شرایط مداخله متفاوت از نوزاد متناظر خود در شرایط کنترل هستند. در مورد دوم، افراد در شرایط آزمون و کنترل یکسان هستند، اما زمان گذشته است و تغییرات بسیاری، به غیر از مداخله، برای نوزادها اتفاق افتاده است (از جمله آسیب دائمی وارد آمده بواسطه فنیل‌آلانین در طول سه ماه ابتدای زندگی). بنابراین در هنگام طراحی آزمایش دو وظیفه اصلی وجود دارد. یکی، طراحی یک منبع استنباط بسیار باکیفیت اما الزاماً غیرکامل[۱۹]، و دیگری، درک چگونگی تفاوت و تمایز این منبع مشابه، نسبت به شرایط مداخله.

در واقع، استدلال خلاف‌واقع اساساً امری کیفی است، زیرا استنباط علّی -حتی در آزمایشها- اساساً کیفی است (Campbell, 1975; Shadish, 1995a; Shadish & Cook, 1999). آماردان‌ها برخی از این نکات را به طور قراردادی در قالب موارد خاصی که برخی اوقات آن را مدل علّی روبین می‌نامند (Holland, 1986; Rubin, 1974, 1977, 1978, 1986)، توضیح می‌دهند[۲۰]. این کتاب درباره آمار نیست و ما به شرح جزئیات مدل روبین نخواهیم پرداخت (برای آشنایی با این مدل به مطالعه West, Biesanz & Pits, 2000 رجوع نمایید). تأکید اصلی مدل روبین تحلیل علّت در آزمایش‌هاست؛ و مفروضات بنیادین و پایه‌ای آن با مفروضات این کتاب همخوانی دارد. مدل روبین به طور وسیعی در مطالعات مورد-کنترل در بهداشت عمومی و پزشکی (Holland & Rubin, 1988)، در تحلیل مسیر در جامعه شناسی (Holland, 1986)، و در پارادکسی که لُرد (Lord, 1967) آن را در روانشناسی معرفی کرد و بعدها منشاء نواوری‌های آماری فراوانی قرار گرفت، بکار گرفته شده است (Dawid, 2000; Pearl, 2000). روشن است که مدل روبین مدلی بسیار عمومی، با کاربردهای روشن و مستحکم است. محققین و دانشجویان باید این مدل و نقدهای وارد آمده به آن را برای انجام بهتر مطالعات علّت‌یابی مطالعه نمایند.

رابطه علّی

از کجا و چطور می‌توان فهمید که علّت و اثر به یکدیگر مربوط هستند؟ بر اساس تحلیل کلاسیکی که جان استوارت میل (فیلسوف قرن هفدهم) پیشنهاد می‌کند، یک رابطه علّی زمانی وجود خواهد داشت که (۱) علّت قبل از اثر اتفاق بیافتد، (۲) علّت به اثر مرتبط باشد، (۳) نتوان توضیح جایگزین منطقی دیگری برای اثر مشاهده شده (بغیر از علّت مورد بحث) پیدا کرد. این سه مشخصه منعکس‌کننده آن چیزی است که در آزمایش‌ها رخ می‌دهد. در یک آزمایش، (۱) علّت مفروض را دستکاری می‌کنیم و پس از آن برونداد (نتایج) را مشاهده می‌کنیم؛ (۲) بررسی می‌کنیم که آیا واریانس موجود در علّت با واریانس موجود در اثر مرتبط است؛ و (۳) روش‌های متفاوتی را در جریان کار آزمایش بکار می‌گیریم تا موجّه بودن دیگر توضیحات ممکن برای اثر

19 imperfect
20 formalized

مشاهده شده را کاهش دهیم. بعلاوه، روش‌های دیگری برای پی بردن به موجّه بودن دیگر توضیحات ممکنی که قادر به بی‌اثر کردن[۲۱] آنها نیستیم را به خدمت می‌گیریم (بخش اعظمی از این کتاب به این روش‌ها اختصاص دارد).

از این رو آزمایش‌ها به خوبی برای مطالعه روابط علّی آماده و تجهیز شده‌اند. معمولاً هیچ روش علّی دیگری به این خوبی با مشخصات روابط علّی تناسب ندارد. تحلیل مایلز به نقاط ضعف دیگر روش‌ها اشاره می‌کند. برای مثال، در بسیاری از مطالعات همبستگی، فهمیدن اینکه کدامیک از متغیرها اول رخ داده غیرممکن است، در نتیجه تعریف یک رابطه علّی میان آنها متزلزل و غیرمطمئن است. درک منطق روابط علّی و چگونگی تعریف اجزاء و اصطلاحات اصلی این روابط (علّت، اثر، ...)، به محققین در نقد مطالعات علّت‌یابی کمک می‌کند.

علّیت، همبستگی و متغیرهای کمکی[۲۲]

یکی از اصول شناخته شده تحقیق آن است که «همبستگی اثبات‌کننده علّیت نیست». علّت این امر آن است که ما نمی‌دانیم کدام متغیر اول آمده، و همچنین نمی‌دانیم آیا تبیینهای جایگزین دیگری برای اثر مفروض وجود دارد یا خیر. برای مثال، فرض کنید سطح تحصیلات و درآمد با یکدیگر همبستگی دارند. آیا لازم است برای پرداخت هزینه‌های تحصیل درآمد بالایی داشته باشید؟ یا اینکه اول شما باید تحصیلات بهتری کسب کنید تا بتوانید کاری با درآمد بالا پیدا کنید. هر کدام از این دو امکان می‌تواند درست باشد. در نتیجه هر دو این سؤالها نیازمند بررسی هستند. تا زمانی که این بررسیها انجام نشده و توسط محققین ارزیابی نشده باشد، یک همبستگی ساده نمی‌تواند نشان دهد کدام متغیر اول بوده است. گذشته از آن، همبستگی‌ها قابلیت اندکی برای بی‌اثر کردن توضیحات جایگزین در مورد رابطه میان درآمد و تحصیلات دارند. رابطه می‌تواند از اساس علّی نبوده باشد، و متغیر سومی (که عموماً متغیری کمکی خوانده می‌شود) مسئول اثرات مشاهده شده باشد. مثلاً هوش یا موقعیت اجتماعی-اقتصادی خانواده فرد باعث تحصیلات عالی و درآمد بالا شده است. اگر هوش بالا موجب موفقیت در تحصیل و شغل شده باشد، پس باهوشها میان تحصیلات و درآمد همبستگی برقرار می‌کنند، نه چون تحصیلات موجب درآمد می‌شود (و یا برعکس)، بلکه چون هردو آنها (تحصیلات و درآمد) به علت هوش بوجود آمده‌اند. بنابراین یکی از وظایف اصلی در مطالعات آزمایشی، مشخص کردن انواع مختلف متغیرهای مزاحمی است که می‌توانند در زمینه‌ای خاص عمل نمایند؛ و همچنین، درک نقاط قوت و ضعف روش‌های مختلفی که از آن طریق می‌توان با اثر این متغیرهای کمکی مقابله کرد.

21 Rule out
22 confounds

دلایل قابل‌دستکاری و غیرقابل‌دستکاری

با درک عمومی از مفهوم آزمایش، منطقی است اگر بگوییم «اجازه بدهید ببینم اگر مستمری بگیران را مجبور به اشتغال کنیم چه اتفاقی می‌افتد؟»؛ اما منطقی نیست اگر بگوییم « ببینیم با تغییر مردی بالغ به دختربچه‌ای سه ساله چه اتفاقی می‌افتد؟». همین منطق در آزمایش‌های علمی نیز مصداق دارد. آزمایش‌ها اثرات متغیرها یا چیزهایی را کشف می‌کنند که قابل‌دستکاری (بالا یا پایین کردن عمدی) است؛ مثل دوز داروها، مقدار پرداخت مستمری، نوع بیکاری، نوع یا مقدار جلسات رواندرمانی، یا تعداد دانش‌آموزان داخل یک کلاس. مشخصات و یا رویدادهای غیرقابل‌دستکاری، مانند انفجار یک ابرستاره و یا سن، جنسیت بیولوژیک و مشخصات ژنتیکی افراد را نمی‌توان به عنوان علّیت در آزمایش در نظر گرفت، چون نمی‌توانیم آنها را به طور حساب‌شده تغییر داده تا ببینیم چه اتفاقی رخ خواهد داد. در نتیجه، اغلب دانشمندان و فلاسفه با این مسأله موافقند که کشف و یافتن اثرات علّت‌های غیرقابل‌دستکاری بسیار دشوارتر است.

بحث بر سر این نیست که تمامی‌علّت‌ها باید قابل دستکاری باشند، بلکه علّت‌های آزمایشی باید چنین خصوصیتی داشته باشند. بسیاری از متغیرهایی که آنها را به عنوان علّت می‌شناسیم به طور مستقیم قابل‌دستکاری نیستند. در نتیجه، اینکه یک نقص ژنتیکی منجر به PKU می‌شود، کاملاً پذیرفته شده است، با وجود آنکه این علّت قابل‌دستکاری نیست. می‌توانیم این گونه علّت‌ها را به طور غیرمستقیم و در مطالعات غیرآزمایشی، و یا حتی در آزمایش‌ها و از طریق فرایندهای دستکاری ژنتیکی که طی آن می‌توان از عمل و یا اثر یک ژن جلوگیری کرد (مانند استفاده از رژیم غذایی برای از کار انداختن اثرات بیولوژیکی ژن) بررسی کنیم. هم ژن غیرقابل‌دستکاری و هم رژیم قابل‌دستکاری را می‌توان بعنوان علّت در نظر گرفت. هردو با عقب ماندگی ذهنی ناشی از PKU کوواریانس دارند، هر دو پیش از عقب‌ماندگی ذهنی رخ می‌دهند، و امکان یافتن توضیحات جایگزین برای اثر ژن و یا رژیم بر روی عملکرد شناختی وجود دارد. با این وجود، بررسی اثر رژیم غذایی به عنوان یک عامل قابل‌دستکاری، دو مزیت نسبت به بررسی اثر ژن غیرقابل دستکاری دارد. اول اینکه، تنها رژیم می‌تواند راه‌حلّی مستقیم برای مقابله با مشکل در اختیار ما قرار دهد. دوم اینکه، خواهیم دید که مطالعه متغیرهای قابل‌دستکاری استنباطهای خلاف‌واقع مرغوبتری (از طریق روش‌هایی مانند تخصیص تصادفی) تولید می‌کنند. هنگامی‌که افراد دچار نقایص ژنتیکی غیرقابل‌دستکاری، با افرادی که واجد این نقایص نسیتند مقایسه می‌شوند، این دو دسته ممکن است از بسیاری جهات (به غیر از جنبه نقص ژنتیکی) با یکدیگر تفاوت داشته باشند. بنابراین ساختن استنباط خلاف‌واقع در مورد اینکه برای افراد دارای PKU ژنتیکی چه اتفاقی می‌توانست رخ دهد، بسیار دشوارتر خواهد بود.

با این وجود، علّت‌های غیرقابل دستکاری نیز می‌بایست با هر ابزار در دسترس و مفید ممکن مورد بررسی قرار بگیرند. زیرا این علّت‌ها می‌توانند کمک کنند تا بتوان عوامل قابل‌دستکاری واجد قابلیت حل مشکل موردنظر را

پیدا کنیم. مثال PKU به خوبی نشان‌دهنده این مطلب است. محققین پزشکی از ابتدا متوجه این موضوع نشده بودند که می‌توان با رژیم غذایی از عقب‌ماندگی ذهنی کودکان PKU جلوگیری کرد. بلکه آنها ابتدا متوجه مشخصات بیولوژیکی غیرقابل‌دستکاری کودکان عقب‌مانده متأثر از PKU شدند، و دریافتند که سطح فنیل‌آلانین به طور غیرطبیعی بالا بوده، و این مسأله همراه با بروزمشکلات متابولیک و ژنتیکی ویژه‌ای در این کودکان است. این یافته‌ها جهت روش‌های بهبود را نشان داد و باعث شد تا دانشمندان درمان‌هایی که احتمال اثربخشی داشت را مورد آزمایش قرار دهند. در نتیجه بواسطه رشته‌ای از مطالعات با اهداف و اشکال متنوع و درجات متفاوتی از کاهش عدم اطمینان، رژیم جدیدی طراحی شد. برخی از این مطالعات آزمایشی، و برخی دیگر غیرآزمایشی بودند.

علاوه براین، آزمایش‌های معادل[۲۳] (قابل قیاس) را نیز می‌توان برای بررسی علّت‌های غیرقابل‌دستکاری بکارگرفت. در این آزمایش‌ها عاملی که معادل علّت مورد نظر است دستکاری می‌شود. مثلاً نمی‌توانیم نژاد کسی را تغییر دهیم، اما می‌توان در افراد داوطلب به طور شیمیایی تغییرات رنگدانه‌سازی پوستی ایجاد کرد (اگرچه این تغییرات دقیقاً معادل اینکه همواره و برای تمام عمر سیاه‌پوست باشی نیست). به همین طریق، رویدادهای واقع شده در گذشته – که طبیعتاً غیرقابل دستکاری هستند- می‌توانند نوعی آزمایش طبیعی را -که حتی تصادفی‌سازی نیز شده باشد- شکل دهند؛ مانند زمانی که سیستم قرعه‌کشی برای انتخاب سربازان اعزامی به جنگ ویتنام، نتایج متفاوتی را مورد بررسی قرار می‌داد (Angrist, Imbens, & Rubin, 1996a; Notz, Staw, & Cook, 1971).

اگرچه آزمایش بر روی علّت‌های قابل دستکاری، کار کشف اثرات این علّت‌ها را آسانتر می‌سازد، اما آزمایش‌ها ابزار کامل و بی‌نقصی برای بررسی علّت‌ها نیستند. برخی اوقات در آزمایش‌ها شرایطی که آزمون در آن اتفاق می‌افتد به گونه ای شکل می‌گیرد (طراحی می‌شود) که تناسب میان این شرایط و موقعیتی که نتایج باید به آن تعمیم داده شود، کاهش می‌یابد. و یا اطلاعات موجود در خصوص اثر علّت‌های قابل‌دستکاری هیچ دانشی نسبت به نحوه و چرایی رخ دادن این اثرات بدست نمی‌دهد. مثلاً چه سؤالاتی ارزش پرسیدن دارد؟ نیاز به مداخله یا درمان چقدر است؟ نحوه توزیع یک علّت در میان جامعه چگونه است؟ آیا درمان مورد نظر با پایبندی به اصول نظری اجرا شده است؟ و در نهایت، باید تا چه اندازه برای نتایج بدست‌آمده ارزش قائل شد؟

بعلاوه، در آزمایش‌ها، ابتدا یک مداخله یا درمان را اجرا می‌کنیم، و تنها پس از دستکاری است که می‌توان اثرات آن را مشاهده و اندازه گیری کرد. اما در برخی دیگر از مطالعات، ابتدا یک اثر مشاهده می‌شود (مانند ایدز)، و سپس به دنبال علّت‌های آن می‌گردیم، خواه این علّت‌ها قابل‌دستکاری باشند، خواه نباشند. آزمایش‌ها در جریان این جستجو به کمک ما نمی‌آیند. اسکریوان (Scriven, 1976) این گونه تحقیقات را به عملکرد یک کارآگاه تشبیه

23 Analogue

می‌کند، که با جرمی مواجه است که به وقوع پیوسته (مثلاً یک دزدی)، کارآگاه الگوی خاصی از شواهد اطراف جرم را مشاهده می‌کند (دزد یک کلاه بسکتبال و یک ژاکت خاص پوشیده و نوع خاصی از اسلحه را بکار برده است)، و سپس به دنبال مجرمینی می‌گردد که روش شناخته‌شده عملکرد آنها[۲۴] با الگوی شواهد موردنظر متناسب باشد؛ و آنها را به عنوان یک مظنون مورد بررسی بیشتر قرار خواهد داد (Ahlbom & Norell, 1990). متخصصین بیماری‌های همه‌گیر[۲۵] روش مشابهی را با عنوان طرح مورد-کنترل[۲۶] مورد استفاده قرار می‌دهند. در این روش، این محققین نتایج سلامتی-بهداشتی خاصی را که در گروه دیگر دیده نمی‌شود، مورد مشاهده قرار می‌دهند و سپس تلاش می‌کنند تا علّت‌های مرتبط با آن را تشخیص دهند (مثلاً افزایش استفاده از تلفن همراه). با همه این احوال، آزمایش‌ها برای پاسخ دادن به تمامی‌انواع سوالات علّی مدنظر دانشمندان علوم اجتماعی کاربرد و ظرفیت ندارند.

توصیف علّی[۲۷] و توضیح علّی[۲۸]

نقطه قوت منحصربه‌فرد آزمایشها، توانایی توصیف نتایج منتسب به مداخله در حال تغییر است. این قابلیت را توصیف علّی می‌گویند. البته آزمایش‌ها توانایی چندانی برای تبیین و توضیح مکانیسمی‌که بواسطه آن، و شرایطی که تحت آن، رابطه علّی مورد نظر وجود خواهد داشت را ندارند – یعنی همان چیزی که از آن به عنوان توضیح علّی یاد می‌شود. برای مثال، اغلب کودکان بسیار سریع توصیف علّی رابطه میان فشار دادن کلید برق و روشن شدن لامپ و در نتیجه اتاق را می‌آموزند. اگرچه معدودی از آن‌ها (و حتی بالغین) می‌توانند به طور کامل توضیح دهند که چرا این چراغ روشن می‌شود. برای اینکه فردی بتواند این کار را انجام دهد، باید بتواند مداخله موردنظر را (عمل فشار دادن کلید روشنایی) به اجزاء علّی موثر (مانند بسته شدن یک مدار الکتریکی) و اجزاء غیرضروری آن (مانند اینکه آیا کلید موردنظر با دست فشرده شود یا یک حسگر حرکتی) تجزیه کند. همچنین فرد باید بتواند همین کار را برای اثر نیز انجام دهد. چراغ می‌تواند رشته‌ای یا فلورسنت باشد، اما همچنان نور تولید می‌کند. پس برای ارائه توضیح کامل، فرد باید بتواند نشان دهد اجزاء علّی موثر مداخله چطور بواسطه فرایندهای واسطه‌ای (مانند عبور الکتریسیته از مدار و برانگیختگی فوتون‌ها) بر روی عناصر تأثیرپذیرفته اثر گذاشته‌اند. واضح است که علّت روشن شدن چراغ، مجموعه پیچیده‌ای از عوامل متعدد است. برای آن دسته از فلاسفه که علّت را مترادف با تشخیص مجموعه متغیرهایی می‌دانند که ضرورتاً و بطور بی‌نقص و اجتناب‌ناپذیری منجر به وقوع اثر می‌شوند، صحبت از علّت امکان‌پذیر نیست، مگر زمانی که تمام عوامل مرتبط با آن شناخته

24 Mondus Operandi
25 Epistemologist
26 Case-control
27 Causal description
28 Causal explanation

شده باشند. از نظر این فلاسفه توصیف علّی در غیاب توضیح علّی بی‌معناست. دلایل فلسفی این دیدگاه هر چه باشد، نمی‌توان از علوم اجتماعی انتظار چندانی برای برآورده کردن انتظار وجود چنین توضیح علّی کاملی را داشت.

اهمیت کاربردی توضیح علّی هنگامی بیشتر جلوه‌گر می‌شود که کلید چراغ نتواند چراغ را روشن کند، و تعویض لامپ (یک دستکاری ساده دیگر) نیز نتواند مشکل را حل کند. اینجا دانش توضیح است که می‌تواند دریافتن راه حل برای برطرف کردن مشکل کمک نماید (مثلاً پیدا کردن و تعمیر کردن بخشی از مدار آسیب‌دیده). یا هنگامی‌که می‌خواهیم مکانی که چراغ ندارد را روشن کنیم، در اختیار داشتن دانش توضیحی باعث می‌شود تا دقیقاً بدانیم کدامیک از اجزاء رابطه علّت و اثر برای ایجاد روشنایی در یک مکان بدون چراغ ضروری هستند، و کدام عناصر غیرمرتبط هستند. توضیح علّی ممکن است این باشد که ضرورت دارد منبعی از الکتریسته وجود داشته باشد، و اینکه منبع می‌تواند فرم‌های ملکولی متعددی از جمله باطری، یک ژنراتور، توربین بادی و یا یک باتری خورشیدی داشته باشد. همچنین باید یک مکانیسم سوئیچ برای بستن مدار وجود داشته باشد که این مکانیسم هم می‌تواند فرم‌های ملکولی متفاوتی از جمله تماس دو سیم لخت، و حتی یک حسگر حرکتی که با ورود یک فرد به اتاق سوئیچ را بگرداند، داشته باشد. بنابراین توضیح علّی راهی مهم برای تعمیم توصیف‌های علّیست؛ چون نشان می‌دهد کدامیک از عناصر علّی باید ضرورتاً به موقعیت‌های دیگر منتقل شوند [برای آنکه بتوان نتایج مشابهی بدست آورد].

این مزیت توضیح علّی نشان‌دهنده اولویت و پرستیژ آن در میان تمامی‌علوم است، و همینطور بیانگر این موضوع است که چرا به مجرد اینکه رابطه علّی مهم و جدیدی کشف می‌شود، هجوم تحقیقات پایه‌ای برای توضیح چرایی و نحوه بروز آن آغاز می‌شود. این تلاش‌ها تجزیه یک علّت به اجزاء اثربخش علّی آن، تجزیه یک اثر به اجزاء متأثر شده (علّی) آن، و شناسایی فرایندهایی که از آن طریق عناصر اثربخش علّی عناصر نتیجه ای را به طور علّی تحت تاثیر قرار می‌دهند، را در بر می‌گیرد.

این مثال‌ها همچنین نشان‌دهنده موازی بودن علّیت توصیفی و علّیت ملکولی و مولی[۲۹] است. تشریح علّی عموماً با روابط ساده دومتغیره میان مداخله‌های مولی و نتایج مولی سروکار دارد. مولی در اینجا به معنای بسته‌ای متشکل از اجزاء متعدد و متنوع است. برای مثال ممکن است دریابیم که رواندرمانی افسردگی را کاهش می‌دهد، این یک رابطه علّی توصیفی ساده میان یک بسته مداخله مولی و یک نتیجه مولی است. اگرچه، رواندرمانی مشتمل بر آیتم‌هایی (گویه‌هایی) است که به عناصر روانشناختی، شناختی و احساسی افسردگی تعلق دارند. علّیت توصیفی این علّت و اثرهای مولی را به اجزاء مولکولی آنها می‌شکند، بنابراین می‌تواند نشان دهد که

29 Molar

تعاملات کلامی و عناصر دارونمایی درمان، هر دو باعث تغییر در علائم شناختی افسردگی می‌شوند، اما پرداخت برای خدمات رواندرمانی چنین اثری ندارد؛ اگرچه این عنصر هم بخشی از بسته مولی مداخله است.

حال اگر آزمایش‌ها نمی‌توانند این دانش را در اختیار ما قرار دهند، چرا تا این اندازه در علم به آنها اهمیت داده می‌شود- علی‌الخصوص در علوم اجتماعی پایه، که در آن نظریه و توضیح علّی بسیار ارزشمند است. پاسخ آن است که دوگانگی میان علّیت توصیفی و توضیحی در عمل -نسبت به بحثهای انتزاعی درباره علّیت- کمتر روشن و مشخص است. اولاً، بسیاری از کارهای توضیحی علّی شامل زنجیره‌ای از اتصالات و روابط توصیفی علّی هستند که در آن، یک رویداد باعث بروز رویداد بعدی می‌شود. آزمایشها کمک می‌کنند که این رابطه‌ها و اتصالات در هر زنجیره را آزمون کنیم. ثانیاً، آزمایشها کمک می‌کنند تا بتوانیم میان روایی نظریه‌های توضیحی رقیب تمیز قائل شویم (مثلاً از طریق آزمون روابط واسطه‌ای ۳۰ پیشنهاد شده در هر یک از این نظریه‌ها). ثالثاً، برخی آزمایشها به بررسی این موضوع را می‌پردازند که آیا یک رابطه علّی توصیفی، از نظر جهت و قدرت، تحت شرایط A، در مقایسه با شرایط B تغییر می‌کند یا نه (شرایط در اینجا به عنوان متغیری مداخله‌گر۳۱ عمل می‌کند، که می‌تواند شرایطی که تحت آن، اثر مورد نظر وجود خواهد داشت را توضیح دهد). رابعاً، برخی آزمایشها مشاهده‌هایی کمّی و کیفی در مورد ارتباطات و اتصالات موجود در زنجیره توضیح علّی (متغیرهای واسطه‌گر) ارائه می‌کنند، که می‌تواند برای تولید و بررسی توضیحات اثرات توصیفی علّی بکار بیاید. آزمایشها همچنین در زمینه‌های کاربردی علوم اجتماعی که در آن شناسایی راه حلهای کاربردی برای یک مسأله اهمیتی بسیار زیاد داشته، و شاید اولویت بیشتری از توضیح مسأله داشته باشد، بسیار ارزشمند محسوب می‌شود. توضیح دادن همواره برای شناسایی راه‌حل‌های کاربردی ضرورت ندارد. لونتاین (Lewontin, 1997) این مسأله را با مثالی از پروژه ژنوم انسانی مطرح می‌کند. این پروژه یک تحقیق چند میلیارد دلاری برای نقشه‌برداری از ژنوم انسانی است، و امید آن می‌رود که نتایج این پروژه بتواند علّتهای ژنتیکی بیماریها را روشن سازد. اما لونتاین نسبت به جنبه‌های مختلف این پروژه چندان خوشبین نیست. وی می‌گوید:

«آنچه در این پروژه رخ می‌دهد تفاوت میان توضیح و مداخله است. بسیاری از بیماری‌ها را می‌توان بواسطه عدم‌توانایی یک ارگان برای تولید یک پروتئین طبیعی توضیح داد، ناتوانی‌ای که نتیجه یک جهش ژنیست. اما مداخله نیازمند آن است که پروتئین نرمال در جای درست، در سلولهای درست، در زمان درست، و به مقدار درست عرضه شود، تا منجر به عملکرد طبیعی سلول شود. آنچه کار را دشوارتر می‌کند، آن است که شاید لازم باشد پروتئینهای غیرنرمال در زمانی حساس و خاص از سلول دور نگه داشته شوند. هیچکدام از این اهداف با دانستن توالی DNA ژن آسیب دیده بدست نمی‌آید (ص ۲۹)».

30 Mediating
31 Moderator

کاربردهای علمی و پیشرفت‌های نظری یکباره نمایان نمی‌شوند بلکه شاید دهه‌ها کار و مطالعات پیگیری۳۲ لازم باشد تا بتوان نتایج آنها را به طور ملموس دید. از جمله اینگونه مطالعات پیگیری، می‌تواند آزمون روابط علّی ساده توصیفی باشد. همین مسأله در مورد مثال داروی اندواستاتین که پیش از این به آن اشاره شد، قابل مشاهده بود. دانشمندان می‌دانستند که دارو از طریق قطع عرضه خون به تومورهای سرطانی کار میکند، اما برای آنکه بتوانند بطور موفقیت‌آمیزی دارو را مورد استفاده قرار دهند، لازم بود تا دارو را در محل، بُعد و عمق درست تزریق کنند. جزئیاتی که به عنوان بخشی از توضیح علّی اثرات دارو به حساب نمی‌آمد.

در نهایت، توصیفها و توضیح‌های علّی در یک موازنه (بده‌بستان) حساس قرار دارند. آنچه که آزمایشها به نحو احسن از پس انجام آن بر می‌آیند، ارتقاء توصیفهای علّی است. اما در توضیح علّی به این خوبی عمل نمی‌کنند. اگرچه اغلب آزمایش‌ها را می‌توان به گونه‌ای طراحی کرد که بتوانند بهتر از آنچه معمولاً انجام می‌دهند، توضیحات علّی ارائه نمایند. علاوه بر این، در ارائه توصیفهای علّی نیز آزمایشها اغلب رویدادهای مولی را مورد بررسی قرار می‌دهند. این رویدادها در مقایسه با فرایندهای ملکولی واسطه‌ای، احتمالاً ارتباط ضعیفتری با نتایج دارند؛ علی‌الخصوص نسبت به فرایندهایی که در زنجیره توضیح علّی به بروندادها نزدیکتر هستند. با وجود این، بسیاری از توصیفهای علّی قابل اتکاء بوده، و از میزان استحکام لازم برای اینکه بلوکهای سازنده‌ی سیاستها و نظریه‌های جدید را تشکیل دهند برخوردارند. به عنوان نمونه روایی ادعاهایی که در ادامه می‌آیند را در نظر بگیرید، «جداسازی دانش‌آموزان سیاه و سفید در مدارس منجر به مهاجرت سفیدپوستان در دهه ۱۹۵۰ و ۱۹۶۰ از مناطق مختلط با سیاهپوستان به مناطق یکدست سفیدپوست در این کشور شد»، «رواندرمانی موجب افزایش سلامت روانی می‌شود»، و با اینکه «رژیم غذایی مناسب عقب‌ماندگی ذهنی ناشی از PKU را کاهش می‌دهد». چنین روابط علّی قابل اتکایی می‌توانند برای سیاستگذاران، دانشمندان و فعالین در صنعت کاربردها و منافع فراوانی داشته باشد.

توصیفهای مدرن از آزمایشها

برخی اصطلاحات مورد استفاده در توصیف آزمایش مدرن (به جدول ۱.۱ نگاه کنید)، منحصر به فرد، روشن و در کاربرد یکدست و منسجم بوده، و برای کاربردهای مختلف مورد استفاده قرار می‌گیرند. مشخصه مشترک تمام آزمایشها کنترل مداخلهست۳۳ (البته خود مداخله نیز می‌تواند فرمها و اشکال متفاوتی داشته باشد). بنابراین به بیان مُستر (Mosteller, 1990, p.225)، «در یک آزمایش محقق نحوه بکارگیری مداخله را کنترل می‌نماید»؛ و به زعم یارمکو و همکارانش (Yaremko, Harari, Harrison, & Lynn, 1986, p.72)، «یک یا بیش از یک متغیر مستقل دستکاری می‌شوند تا اثرات آنها بر یک یا چند متغیر وابسته مورد بررسی قرار گیرد». اگرچه در طول زمان اقسام

32 Follow-up
33 Manipulation

متفاوتی از آزمایشها در پاسخ به نیاز و پیش‌زمینه‌های علوم مختلف شکل گرفته است (Winstone, 1990; Winston & Blais, 1996).

جدول ۱.۱: برخی واژگان آزمایش

آزمایش: مطالعه ای که طی آن، یک مداخله به طور حساب شده‌ای انجام می‌شود تا اثرات آن مشاهده شود.

آزمایش تصادفی: آزمایشی که در آن افراد طی فرایندی تصادفی (مانند استفاده از پرتاب سکه، یا جدول اعداد تصادفی) به شرایط دریافت مداخله یا شرایط عدم‌دریافت آن اختصاص داده می‌شوند.

شبه‌آزمایش: آزمایشی که در آن افراد به صورت تصادفی به شرایط (آزمون و کنترل) تخصیص داده نمی‌شوند.

آزمایش طبیعی: در واقع آزمایش نیست، چون در این مطالعات معمولاً نمی‌توان علّت را دستکاری کرد. مطالعه‌ای که به مقایسه رویدادی که به طور طبیعی اتفاق می‌افتد (مثل یک زلزله)، با حالتی که رویداد مذکور در آن اتفاق نیافتاده می‌پردازد.

مطالعه همبستگی: معمولاً مترادف با مطالعات غیرآزمایشی یا مشاهده‌ای بکار برده می‌شود. مطالعه‌ای که به سادگی تنها اندازه و جهت یک رابطه میان متغیرها را مورد مشاهده قرار می‌دهد.

آزمایش تصادفی[۳۴]

آزمایشهای تصادفی اولین بار در کشاورزی مورد استفاده قرار گرفت. اما بعدها به دیگر رشته‌ها نیز راه یافت. زیرا این روش نویدبخش اعمال کنترل بر منابع فرعی[۳۵] تولید واریانس – بدون نیاز به ایزوله فیزیکی در آزمایشگاه - بود. مشخصه اصلی آزمایش مهم و روشن است. مداخله‌های مختلف را به طور شانسی (برای مثال از طریق پرتاب سکه یا استفاده از جدول اعداد تصادفی) به واحدهای[۳۶] آزمایشی تخصیص می‌دهیم. اگر تخصیص تصادفی به درستی انجام شود، می‌تواند از واحدهایی که به طور احتمالی[۳۷] و به طور متوسط به یکدیگر شبیه هستند، دو یا بیشتر گروه بدست آورد.[۳۸] از این رو، هر گونه تفاوتی که در انتهای مطالعه میان دو گروه مشاهده می‌شود را

34 Random experiments

35 Extraneous

۳۶ - واحدها می‌توانند افراد، حیوانات، بازه‌های زمانی، موسسات، یا تقریباً هر چیز دیگری باشد. معمولاً در آزمایشهای میدانی (field) این واحدها عبارتند از افراد یا مجموعه ای از افراد، مانند کلاسهای درس یا محلهای کار. در این کتاب دو عبارت تخصصی افراد به شرایط و تخصیص شرایط به افراد به یک معنی و به جای یکدیگر بکار برده می‌شود.

37 Probabilistically

۳۸ - کلمه probabilistically بسیار با اهمیت است. این کلمه با جزئیات بیشتر در فصل ۸ توضیح داده می‌شود.

می‌توان در نتیجه وجود مداخله، و نه به دلیل تفاوتهایی که از ابتدای مطالعه (قبل از شروع مطالعه) میان دو گروه وجود داشته قلمداد کرد. بعلاوه، در صورت وجود برخی پیش فرضها، آزمایش تصادفی تخمینی از اندازه اثر مداخله که دارای مشخصات آماری مطلوب است، و همچنین تخمینی از احتمال اینکه اثر درست در محدوده فاصله اطمینان تعریف شده قرار داشته باشد، بدست می‌دهد. این مشخصه‌های آزمایشها به قدری ارزشمند است که در حوزه تحقیقی مانند پزشکی از آزمایش تصادفی اغلب به عنوان طلای استاندارد برای تحقیق نتایج مداخله یا درمان یاد می‌شود[۳۹].

اصطلاح دیگری که به طور مبهم و غیرمنسجم در موقعیتهای مختلف بکار برده می‌شود، و در ارتباطی نزدیک با آزمایش تصادفی قرار دارد، اصطلاح «آزمایش حقیقی» است. در برخی کتابها این اصطلاح مترداف «آزمایش تصادفی» بکار برده شده است (Rosenthal & Rosnow, 1991). برخی دیگر آن را به صورت کلی‌تر و برای اشاره به هر مطالعه‌ای که در آن یک متغیر مستقل به صورت حساب‌شده دستکاری می‌شود (Yaremko et al., 1986)، و یک متغیر وابسته مورد ارزیابی قرار می‌گیرد، بکار برده‌اند. با در نظر گرفتن ابهام موجود در این اصطلاح و اینکه صفت «حقیقی» در اصطلاح «آزمایش حقیقی» این حس را القاء می‌کند که یک روش درست واحد آزمایشی وجود دارد، لذا این کتاب از بکار بردن آن اجتناب می‌کند.

شبه-آزمایش

بخش اعظمی از این کتاب به دسته‌ای از طرحهای مطالعاتی که استنلی و کمپبل (Campbell & Stanley, 1963) از آنها با عنوان شبه‌آزمایش یاد می‌کنند، اختصاص دارد[۴۰]. شبه‌آزمایشها با دیگر انواع آزمایشها از نظر هدف (آزمودن فرضیات علّی توصیفی در خصوص دلیلهای قابل دستکاری)، و بسیاری از دیگر جزئیات ساختاری، مانند حضور مکرر گروههای کنترل و اندازه گیری‌های پیش‌آزمون، که برای پشتیبانی از یک استنباط خلاف‌واقع در مورد اتفاقی که می‌توانست در غیاب مداخله رخ بدهد، طراحی می‌شوند، مشترک هستند. اما شبه‌آزمایشها بنابر تعریف فاقد تخصیص تصادفی است. تخصیص به شرایط متفاوت در مطالعه یا از طریق انتخاب شخصی انجام می‌شود. به این معنی که افرادی که خودشان تمایل داشته باشند برای درمان داوطلب شوند، برای گروه درمان

---

۳۹ - اگرچه اصطلاح آزمایش تصادفی شده پیوسته به این صورت در این کتاب و در بسیاری از زمینه‌ها (field) بکار برده می‌شود، آماردانها اصطلاح تقریباً مشابه آزمایش تصادفی (random experiment) را به صورتی متفاوتی و برای اشاره به آزمایش‌هایی که نتایج آنها را نمی‌توان با قطعیت پیشبینی کرد بکار می‌بنرند.

۴۰ - کمپبل (۱۹۵۷) برای اولین بار آنها را طرحهای اغماضی یا مسامحه ای (compromise) نامید اما خیلی زود این اصطلاح را تغییر داد. روزنباوم (۱۹۹۵a) و کوکران (۱۹۶۵) این طرحها را طرحهای مشاهده ای (observational) می‌نامند،اصطلاحی که در این کتاب از بکار بردن آن اجتناب می‌کنیم چون این اصطلاح مکرراً برای اشاره به مطالعات همبستگی و غیرآزمایشی نیز بکاربرده می‌شود. گرین برگ و شرودر (۱۹۹۷) اصطلاح شبه-آزمایش را برای اشاره به مطالعاتی که در آنها گروهها به طور تصادفی به شرایط مختلف تخصیص داده می‌شوند،بکار می‌برد. اما این کتاب این گونه طرحها را آزمایش‌های تصادفی شده گروهی (group-randomized experiments) می‌نامد.

انتخاب می‌شوند؛ و یا اینکه واحدها توسط محقق یا مجری مطالعه، درمانگر، معلمین، پزشکان و یا دیگران، برای دریافت مداخله انتخاب می‌شوند. اگرچه، محققین در طرحهای شبه‌آزمایشی، همچنان کنترل فراوانی بر انتخاب و برنامه‌ریزی مقیاسها، نحوه انجام تخصیص غیرتصادفی، نوع گروههای مقایسه‌ای[۴۱] (کنترل) که گروه دریافت‌کننده مداخله با آنها مقایسه می‌شود، و دیگر جنبه‌های اجرا و برنامه‌ریزی مداخله دارند. همانطور که کمپبل و استنلی اشاره می‌کنند:

«موقعیتهای طبیعی و اجتماعی فراوانی وجود دارد که در آن محقق می‌تواند چیزی شبیه طرح آزمایشی را بر روی فرایند جمع‌آوری داده مطالعه خود پیاده کند (مانند زمان و افراد مورد اندازه گیری)، حتی زمانی که محقق کنترل کاملی روی برنامه‌ریزی و زمانبندی محرک آزمایشی (زمان مواجهه و اینکه چه کسی در معرض مداخله واقع شود و توانایی تصادفی کردن مواجهه‌ها)، یعنی آنچه یک آزمایش واقعی را ممکن می‌سازد، ندارد. این شرایط در مجموع شمایی از طرحهای شبه‌آزمایشی را به تصویر می‌کشد (Campbell & Stanley, 1963, p. 34)»

در شبه‌آزمایشها، علّت قابل‌دستکاری بوده، و قبل از آنکه اثر محاسبه شود رخ می‌دهد. اگرچه مشخصه‌های طرحهای شبه‌آزمایشی باعث می‌شود تا نتوانند پشتوانه لازم برای استنباطهای خلاف‌واقع فراهم آورند. برای مثال، در طرحهای شبه‌آزمایشی ممکن است گروههای کنترل به طور سیستماتیک با گروه مداخله تفاوت داشته باشند (یعنی به دلیلی غیر از مداخله تفاوت دارند). بسیای از این دلایل خود می‌توانند تبیین یا توضیحی جایگزین برای اثر مشاهده شده به حساب بیایند. بنابراین محققین برای بدست آوردن تخمینی با روایی مناسب از اثر مداخله، باید بتوانند این عوامل را بی‌اثر نمایند (اثر آنها را در تحقیق از میان ببرند). در مقابل، هنگامی‌که تخصیص تصادفی وجود داشته باشد، لازم نیست محقق تا این اندازه نگران توضیحات و تبیین‌های جایگزین باشد. اگر تخصیص تصادفی بدرستی انجام شود، احتمال اثربخشی اغلب جایگزینها از ابتدای مطالعه از میان می‌رود.

در شبه‌آزمایشها، محقق باید توضیح‌ها و تبیین‌های جایگزین برای توضیح در دست بررسی را برشمرده، و تصمیم بگیرد که کدامیک موجّه بوده و لازم است تا مورد توجه قرار گیرد. سپس باید منطق، طراحی، و مقیاسها را در خدمت گرفته و این مسأله را ارزیابی نماید که آیا هر کدام از این تبیین‌های جایگزین در جهتی که توضیح دهنده اثر مشاهده‌شده باشند، عمل می‌کند یا نه. مشکل آنجاست که این تبیینهای جایگزین، همگی از ابتدا قابل‌شناسایی نیستند، و بنابراین نمی‌توان تمامی‌آنها را از ابتدا در نظر گرفت. گذشته از آن، برخی از این تبیین‌ها مختص زمینه خاص مورد مطالعه هستند، و روشهای لازم برای حذف و بی‌اثر کردن آنها به ازای هر گزینه و برای هر زمینه در دست مطالعه، متفاوت است. برای مثال، فرض کنید دو گروه از کودکان که به طور

41 Comparison group

غیرتصادفی دسته‌بندی شده‌اند را در دست مطالعه داریم، یکی از گروهها داوطلب دریافت مداخله شیوه جدید مطالعه است، و دیگری گروه کنترلیست که مداخله موردنظر را دریافت نمی‌کند. اگر گروه مداخله عملکرد بهتری داشته باشد، آیا این برتری در عملکرد به دلیل مداخله بوده است، و یا اینکه به دلیل رشد شناختی بهتری بوده که گروه داوطلب پیش از شروع مداخله نیز از آن برخوردار بوده است؟ (توجه داشته باشید که در یک آزمایش تصادفی‌سازی شده نرخ رشد بلوغ شناختی[۴۲] احتمالاً در میان گروهها برابر است). برای ارزیابی این گزینه جایگزین، محقق باید پیش‌آزمونهای متعددی را به مطالعه اضافه کند، تا روند بلوغ شناختی هر یک از گروهها را پیش از شروع مداخله ارزیابی، و سپس روند مذکور را با روند شکل گرفته بعد از انجام مداخله مقایسه نماید.

توضیح جایگزین دیگر برای اثر مشاهده (برتری گروه آزمون) آن است که گروه کنترلی که بصورت غیرتصادفی انتخاب‌شده، ممکن شامل تعداد بیشتری از کودکان متعلق به طبقه محروم جامعه باشد. این کودکان طبیعتاً دسترسی محدودتری به کتاب در خانه‌هاشان داشته، و احتمالاً والدینی داشته‌اند که کمتر برای آنها کتاب خوانده‌اند (در یک آزمایش تصادفی، هر دو گروه آزمون و کنترل سهم برابری از این طور دانش‌آموزان خواهند داشت).

برای ارزیابی این گزینه، محقق می‌تواند تعداد کتابهای موجود در خانه را محاسبه کرده، مقدار زمان صرف شده توسط والدین برای خواندن کتاب برای کودکان را اندازه گیری کرده، و شاید تعداد دفعات رفتن به کتابخانه را محاسبه نمایند. پس از آن محقق باید ببیند آیا اندازه این متغیرها در میان گروه کنترل و آزمون تفاوت معناداری دارد؟ تفاوتی که بتواند برتری مشاهده‌شده در گروه آزمون را توضیح دهد. واضح است که با افزایش تعداد تبیین‌های موجّه جایگزین، طراحی شبه‌آزمایش دشوارتر و پیچیده تر می‌شود، علی‌الخصوص با در نظر گرفتن این مطلب که هیچگاه نمی‌توان مطمئن بود که تمام تبیین‌های جایگزین را شناسایی کرده‌ایم. کم‌کم به نظر می‌رسد تلاش‌های محقق برای کنترل کردن تبیین‌های جایگزین شبیه تلاش برای بستن زخمی است که در صورت استفاده از تخصیص تصادفی (از ابتدای کار) به این وخامت نمی‌رسید.

تلاش برای بی‌اثر کردن متغیرهای جایگزین، و خارج کردن اثر فرضیات مرتبط با آنها از مطالعه، در راستای منطق ابطال‌پذیری[۴۳] پوپر (Popper, 1959) است. بر اساس این منطق، اطمینان حاصل کردن از اینکه یک نتیجه‌گیری کلی (مثل اینکه همه قوها سفید هستند)، که بر مبانی مجموعه محدودی از مشاهدات بدست آمده (مثلاً تمام قوهایی که مورد مشاهده قرار گرفته‌اند سفید بوده‌اند)، بسیار دشوار است. از دیدگاه پوپر مشاهده یک مورد مخالف فرضیه (مثل دیدن یک قوی سیاه) کافیست تا این نتیجه کلی که همه قوها سفید هستند را ابطال کنیم. پوپر به دانشمندان توصیه اکید می‌کند بجای جستجوی صرف اطلاعاتی که تأیید فرضیات آنهاست، تلاش

42 Maturation
43 Falsification

خود را مصروف ابطال نتایج مورد علاقه خود کنند. نتایجی که در برابر تلاشهای ابطال همچنان دوام می‌آورند، در مجلات و کتابهای علمی‌حفظ می‌شوند، و با آنها به صورت نتایجی موجّه رفتار می‌شود، البته تا زمانی که شواهد بهتری پیدا شود. انجام شبه‌آزمایش یک فرایند ابطال است؛ از آن جهت که باعث می‌شود تا محققین انجام‌دهنده یک آزمایش، یک ادعای کلی را شناسایی کرده، و سپس گزینه‌های جایگزین موجّه و قابل‌اعتنایی که قابلیت ابطال ادعای مذکور را دارند را تولید و بررسی کنند.

اگرچه، چنین تلاشهایی برای ابطال، هیچگاه با آن قطعیتی که مدنظر پوپر بود، انجام نمی‌شود. کوهن (Kuhn, 1962) بر این باور است که ابطال به دو پیشفرض وابسته است، که هیچگاه نمی‌توان آنها را به طور کامل آزمون کرد (از وجود آنها اطمینان حاصل کرد). اولی، اینکه ادعای علّی به طور تام و کامل تعیین و مشخص‌شده[۴۴] است. اما هیچگاه چنین چیزی اتفاق نمی‌افتد، و بسیاری از عناصر و مشخصه‌های ادعای علّی و آزمون آن ادعا قابل‌بحث هستند. برای مثال، کدام نتیجه مطلوب یا مورد توجه است؟ چطور اندازه‌گیری می‌شود؟ شرایط مداخله چیست؟ چه کسی به مداخله نیاز دارد؟ و بسیاری دیگر از تصمیم‌هایی که محققین باید در جریان آزمون روابط علّی اتخاذ کنند. در نتیجه، تأیید نشدن فرضیه‌ها اغلب باعث می‌شود تا نظریه‌پردازان بخشی از نظریه‌های علّی خود را بازتعریف کنند. مثلاً ممکن است محقق شرایط بدیع و جدیدی را تعیین کنند، شرایطی که نظریه آنها تنها در صورت وجود این شرایط صدق می‌کند. این شرایط را ممکن است با در نظر گرفتن مشاهدات نقض‌کننده فرضیات خود بدست آورده باشند. دومی، ابطال‌پذیری نیازمند مقیاسهایی است که به طور کامل و تام انعکاسی از نظریات در حال آزمون باشند. اگرچه، اغلب فلاسفه به این قائل هستند که تمامی‌مشاهدات نظریه-بار[۴۵] یا زیر بار نظریه هستند. مشاهدات از یک سو زیر بار تفاوتهای ظریف موجود در درک افراد و گروههای طراحی‌کننده آزمونها از نظریه هستند. و از سوی دیگر، زیر بار آرزوها، انگیزه‌ها و پیش‌فرضهای مشترک فرهنگی و البته فراعلمی محقق هستند. اگر مقیاسها مستقل از نظریه‌ها نباشند، چطور قادرند آزمونهای مستقلی برای آزمودن نظریه‌ها ارائه کنند. اگر امکان وجود مشاهدات فارغ از نظریه[۴۶] وجود نداشته باشد، متعاقباً امکان وجود دانش قطعی نسبت به اینکه چه چیزی تایید یا ردکننده یک ادعای علّی است، نیز منتفی خواهد بود.

با اینهمه، همچنان می‌توان ویرایشی (نسخه‌ای) خطاپذیر[۴۷] از ابطال را امکان‌پذیر دانست. از منظر رویکرد ابطال خطاپذیر، مطالعه فرضیات علّی همچنان می‌تواند درک ما را نسبت به روندهای کلی و عمومی افزایش دهد؛ علی‌رغم آنکه نمی‌تواند تمامی اقتضائات مرتبط با هر یک از این روندها را لحاظ کند. این رویکرد قائل به

44Specified

45 Theory-laden (توضیح مترجم: اصطلاحی در فلسفه بدین معنی که مشاهده علمی‌با نظریات پس-زیمنه فکری مشاهده گر ارتباط دارد)

46 Theory-neutral

47 Fallibilist

سودمندی و مفید بودن مطالعات علّی است، حتی اگر مجبور باشیم برای انجام این مطالعات، و برای لحاظ کردن اقتضائات جدید و فهم تازه از نظریه، بارها و بارها فرضیات اولیه را بازتعریف کنیم. گذشته از آن، بازتعریف‌های موردنیاز معمولاً جزئی بوده، و ندرتاً مجبور به کنار گذاشتن روندهای کلی و جایگزین کردن آنها با روندها و رویکردهای متضاد هستیم. ابطال‌پذیری همچنین فرض را بر این می‌گذارد که اگرچه مشاهده فارغ از نظریه غیرممکن است، اما مشاهدات می‌توانند به مرز موفقیت حقیقت‌گونه[۴۸]بودن نزدیک شوند، اگر، (۱) به طور مکرر و در طول نظریه‌های متفاوت و در جریان سازه‌پردازی‌های متفاوت، (۲) در طول انواع مختلف [مقیاسهای] محاسبه و اندازه گیری، و در (۳) در زمانهای مختلف همچنان مشاهده شوند. همچنین فرض بر این است که مشاهدات ملهم از چندین نظریه (و نه یک نظریه) هستند؛ در حالی که رویه‌های عملیاتی متفاوت در تحقیق، تحت تأثیر همان نظریه‌ها نیستند. در نتیجه، مشاهداتی که، علیرغم نظریه‌های متفاوت مبنای آنها، مکرراً رخ می‌دهند، موقعیت خاص حقیقت‌گونه پیدا می‌کنند؛ حتی اگر هیچگاه نتوان آنها را به عنوان مشاهده مستقل از نظریه پذیرفت. به طور خلاصه، می‌توان گفت ابطال‌پذیری چیزی بیشتر از آن است که ببینیم مشاهدات، یک فرضیه یا پیش‌بینی را تأیید یا رد می‌کنند. بلکه، این دیدگاه به یافتن و قضاوت در مورد ارزش فرضیات فرعی مرتبط با تعیین‌پذیری[۴۹] محدود فرضیات علّی موردآزمون، ناهمگونی نظریه‌ها، دیدگاهها، موقعیتها و زمانهای مستتر در مقیاسهای ارزیابی علّت و اثر، و هرگونه متغیر دیگری که رابطه آنها را تحت تأثیر قرار می‌دهد، می‌پردازد.

بی‌اثر کردن تمامی‌تبیین‌های جایگزین ممکن برای یک رابطه علّی نه ممکن است (به لحاظ شدنی بودن و هزینه انجام کار) و نه مطلوب. بلکه باید تنها روی بی اثر کردن تبیین‌های موجّه تمرکز کرد. این کار باعث می‌شود تا تعداد تبیینهای جایگزین قابل پیگیری باشد، و همچنان بتوان کار را دنبال کرد. چون در غیر این صورت تعداد تبیین‌های ممکن بی نهایت، و در نتیجه کار بی‌پایان خواهد بود. در ضمن بسیاری از تبیینهای جایگزین برای روابط علّی موردنظر هیچ پشتوانه‌ی کاربردی یا تجربی جدی نداشته، و ارزش توجه ندارند. گرچه باید توجه داشت که در نظر گرفتن بی‌پشتوانگی به عنوان یک معیار پذیرش یا رد فرضیه، خود می‌تواند گمراه کننده باشد. برای مثال، برای مدتهای طولانی علّت زخم معده را سبک زندگی (مثلاً استرس) و تولید اضافی اسید معده می‌دانستند. کمتر دانشمندی می‌توانست تصور کند دلیل زخم معده می‌تواند یک باکتری، ویروس یا جرم باشد، چون تصور بر این بود که یک معده پر از اسید هر یاخته‌ای را در خود حل می‌کند. اما در سال ۱۹۸۲ دو محقق استرالیایی به نامهای بری مارشال و روبین وارن (Barry Marshall & Rubin Warren) باکتری فنری‌شکلی که بعدها آن را هلیکوباکتر نامیدند، را در معده افراد مبتلا به زخم معده کشف کردند. با این کشف، فرضیه سابقاً

48 Fact-like
49 Specificity

غیرموجّه تبدیل به فرضیه‌ای موجّه شد. تا سال ۱۹۹۴ موسسات بهداشتی شرکت‌کننده در کنفرانس اجماع در زمینه سلامت، چنین نتیجه‌گیری کردند که هلیکوباکتر علّت اصلی اغلب زخمهای گوارشی است. بنابراین موجّه دانستن فرضیه‌های جایگزین، نه تنها به منطقاً امکانپذیر بودن آنها، که به باورهای جامعه علمی، تجربیات مشترک، و داده‌های عملی وابسته است.

از آنجا که چنین عواملی بسته به زمینه مطالعات[۵۰] متفاوت است، زمینه‌های مهم علمی مختلف، هرکدام تلاش می‌کنند تا مجموعه‌ای از اطلاعات و دانش مرتبط با گزینه‌های واجد شرایط و با اهمیتی که لازم است در جریان مطالعه کنترل شوند را فراهم آورند. برخی حتی پا را فراتر گذاشته، و برای خود روشهای منحصر به فردی برای انجام این کار طراحی می‌کنند. مثلاً در روانشناسی ابتدایی، بکارگیری یک گروه کنترل، همراه با مشاهدات پیش‌آزمون برای کنترل تبیینهای موجّه جایگزین، ابداع شد. تبیین جایگزین این بود که، انجام پیش‌آزمون و تمرین کردن با محتوی سؤالات آن، موجب شکل‌گیری نوعی پیشرفت عملکرد، حتی بدون وجود مداخله می‌شود. پس برای کنترل اثر احتمالی این تبیین جایگزین، یک گروه کنترل اضافی با مشاهدات پیش‌آزمون به مطالعه اضافه شد، تا اثر مورد نظر را کنترل نماید (Coover & Angell, 1907). بنابراین تمرکز بر موجّه‌بودن یک تبیین جایگزین، یک شمشیر دولبه است. از یک سو، می‌تواند طیف گزینه‌هایی که می‌بایست در فرایند انجام شبه‌آزمایش مورد کنترل قرار بگیرند را کاهش دهد. اما از سوی دیگر، می‌تواند باعث شود تا استنباط علّی منتج از این شبه‌آزمایشها آسیب‌پذیر شوند، چون هر زمان ممکن است یکی از گزینه‌هایی که به نظر غیرموجّه می‌رسید، خود به عنوان یک عامل اثرگذار شناسایی و کشف شود.

آزمایش طبیعی

اصطلاح آزمایش طبیعی به معنای تضاد طبیعی میان یک شرایط کنترل و شرایط مداخله است. تضادی که بطور طبیعی رخ داده است. اغلب خود مداخله را نمی‌توان به طور ارادی دستکاری نمود یا تغییر داد؛ مانند زمانی که دانشمندان با رویکردی پس-نگر[۵۱] به دنبال بررسی این سوال بودند که آیا زمین لرزه در کالفرنیا باعث کاهش قیمت ملک شده است؟ (Brunette, 1995; Murdoch, Singh, & Thayer, 1993) با وجود اینکه براحتی می‌توان گزینه‌هایی موجّه در مورد اثرات زمین لرزه ساخت، و حتی از آنها دفاع کرد، زمین لرزه قبل از آنکه مشاهدات مرتبط با قیمت ملک انجام شود اتفاق افتاده، و به سادگی می‌توان دید که آیا این اتفاق به قیمت ملک ارتباطی دارد یا نه. یکی از منابع مفید ساخت استنباط خلاف واقع، می‌تواند مقایسه قیمت املاک در همان منطقه قبل از زمین لرزه، و یا بوسیله مطالعه و مقایسه مناطقی باشد که در همان بازه زمانی زمین لرزه را تجربه نکرده اند. اگر

50 Context-specific
51 Retrospectively

قیمت املاک در مناطق زلزله‌زده پس از وقوع زمین لرزه به شدت کاهش پیدا کرده باشد، اما این اتفاق در منطقه مورد مقایسه (کنترل) در همان بازه زمانی نیافتاده باشد، دشوار می‌توان توضیح جایگزینی برای کاهش قیمت رخ داده پیدا کرد.

آزمایش طبیعی اخیراً در میان اقتصاددانان از اقبال خوبی برخوردار شده است. اقتصاددانان تا قبل از دهه ۱۹۹۰ با قاطعیت فراوان باور داشتند که می‌توانند با استفاده از تعدیلات آماری روی نابرابری‌های موجود میان گروههای کنترل و مداخله، استنباطهای علّی معتبری تولید کنند. اما انجام دو مطالعه بر روی اثرات آموزش حین‌خدمت نشان داد که این تعدیلات موجب تولید تخمینهایی شده بودند که نه تنها به تخمینهای تولید شده در آزمایشهای تصادفی انجام شده در این زمینه نزدیک نبود، بلکه در آزمونهای حساسیت مدل۵۲ نیز پایداری لازم را نداشت (Fraker & Maynard, 1987; Lalonde, 1986). از این رو، در جستجو برای یافتن روشهای جایگزین، بسیاری از اقتصاددانان به انجام آزمایش‌های طبیعی روی آوردند. به عنوان نمونه در مطالعه اقتصادی اثرات رخ داده در بازار کار شهر میامی، هنگامی‌که تعداد زیادی از زندانیان از زندانهای کوبا آزاد شده بودند و اجازه کار در آمریکا را پیدا کردند، از این نوع آزمایش استفاده شد (Card, 1990). اقتصاددانها فرض را بر این گذاشتند که آزادی زندانیان (یا زمانبندی زمین‌لرزه در مثال قبل)، مستقل از فرایندهایی است که به طور معمول نرخ‌های بیکاری را تحت تاثیر قرار می‌دهند (یا قیمت ملک). در ادامه بحث، اعتبار این فرض را بیشتر مورد مداقه قرار خواهیم داد؛ البته تردیدی نیست که صحت این فرض برای محققین از مطلوبیت فراوانی برخوردار است.

طرح‌های غیرآزمایشی

اصطلاحاتی از قبیل طرح‌های همبستگی، طرح‌های منفعل مشاهده‌ای، و طرح‌های غیرآزمایشی به شرایطی اشاره دارند که در آن، علّیت و اثر مفروض شناسایی و محاسبه می‌شوند، اما دیگر عناصر ساختاری آزمایش‌ها در این مطالعات وجود ندارد. تخصیصهای تصادفی بخشی از طرح مطالعه نیست، همچنین هیچ پیش‌آزمون یا گروههای کنترلی که با استفاده از آنها محقق بتواند یک استنباط خلاف حقیقت را بسازد، در این نوع طرح‌های مطالعاتی وجود ندارد. در مقابل، در این مطالعات اتکاء بر محاسبه تبیین‌های (متغیرهای) جایگزین به صورت مجزا، و سپس کنترل کردن آنها به صورت آماری است. در مطالعات مقطعی۵۳، که در آنها تمامی‌داده‌ها در یک مقطع از زمان از پاسخ‌دهندگان جمع‌آوری می‌شود، محقق حتی نمی‌تواند بفهمد که علّت ابتدا اتفاق افتاده یا اثر.

هنگامی‌که اینگونه مطالعات برای اهداف علّی بکار گرفته می‌شوند، عناصر معقول یا غایب طرح تحقیق می‌تواند مسأله‌ساز باشد؛ مگر اینکه اطلاعات و دانش موجود نسبت به تبیین‌ها و تفسیرهای موجّه بسیار زیاد باشد، یا

52 Model sensitivity

53 Cross-sectional

اینکه تبیینهای جایگزین را بتوان به طور معتبری محاسبه کرد، و یا اینکه مدل اصلی تعیین‌شده۵۴ برای تعدیلهای آماری به خوبی تعیین و مشخص شده باشد. تأمین این شرایط در دنیای واقعی بسیار دشوار است، و بنابراین بسیاری از صاحبنظران (در بسیاری از موارد) این طرحها را دارای قابلیت کافی برای بررسی استنباطهای علّی نمی‌دانند.

آزمایشها و تعمیم روابط علّی

نقطه قوت و برتری آزمایش‌ها در توانایی این طرحها برای پرده برداشتن از استنباطهای علّی است. اما نقطه‌ضعف آنها در این است که نمی‌توان دانست روابط علّی شناسایی‌شده، تا چه میزان قابل تعمیم هستند. یکی از جنبه‌های نوآورانه این کتاب، تمرکز آن بر مسأله تعمیم است. در این قسمت مروری گذرا بر مسائل کلی مرتبط با تعمیم داشته، و بسط این مسائل را به فصول آینده موکول می‌کنیم.

اغلب آزمایشها به میزان زیادی محدود به یک محل هستند۵۵اما اهداف کلی دارند

اغلب آزمایشها تا حد زیادی محدود به یک محل [جغرافیایی] خاص بوده، و برای موضوع خاصی طراحی شده‌اند۵۶. آنها تقریباً همواره با طیف محدودی از مختصات۵۷ (اغلب با تنها یک مختصات)، و با ویرایش خاصی از یک نوع مداخله (و نه نمونه‌ای از تمام ویرایش‌های ممکن) انجام می‌شوند. آزمایشها معمولاًچندین مقیاس را در برمی‌گیرند که هر کدام از این مقیاسها پیش‌فرضهای نظری خاصی دارند، پیش‌فرضهایی که احتمالاً متفاوت از پیش‌فرضهای مبنایی دیگر مقیاسهاست، اما همچنان فاصله بسیار زیادی تا ایده‌ال (یعنی داشتن مجموعه‌ای کامل از تمامی مقیاسهای ممکن) دارند. هر آزمایش به جای آنکه نمونه‌ای را انتخاب نماید که به بهترین نحو نماینده جمعیت تعریف شده باشد، تقریباً همواره نمونه‌ای در دسترس از افراد را برای آزمایش به خدمت می‌گیرد. و در نهایت، آزمایشها به طور اجتناب‌ناپذیری در یک نقطه خاصی از زمان انجام می‌شوند؛ نقطه‌ای که خیلی زود به تاریخ (گذشته) تبدیل می‌شود.

علیرغم این موارد، خوانندگان نتایج آزمایشها، ندرتاً علاقمند به دانستن چیزی راجع به آن زمان خاص یا آن محدوده جغرافیایی خاص هستند. بلکه هدف آنها از مطالعه عموماً دانستن چیزی در مورد سازه‌های نظری موردنظر، و یا سیاستهای بزرگتر است. نظریه‌پردازان اغلب مایلند نتایج بدست آمده از آزمایشها را به نظریاتی که کاربرد مفهومی وسیعی دارند، پیوند بزنند. این کار بیشتر نیازمند تعمیم در (زمینه یا) سطح کلامی۵۸ سازه است،

54 Specified
55 local
56 particularistic
57 Seng
58 Linguistics

تا تعمیم در (زمینه یا) سطح عملیاتی[۵۹] که در یک آزمایش برای نشان دادن این سازه‌ها بکار گرفته شده است. محققین تقریباً همواره به تعمیم نتایج به افراد و شرایطی بیشتر و وسیعتر از آن چیزی که در یک آزمایش بکار گرفته شده است، علاقمند هستند. قطعاً ارزش یک نظریه به وسعت طیف پدیده‌هایی که آن نظریه پوشش می‌دهد وابسته است. به همین ترتیب، سیاستگذاران کنجکاوند بدانند که رابطه علّی مشاهده شده (به طور احتمالی) در مکانهای متعددی که سیاست مورد نظر در آنها به اجرا درخواهد آمد مصداق دارند یا نه. استنباطی که نیازمند تعمیم نتایج به حوزه‌ای فراتر از زمینه مطالعاتی اولیه آزمایش است. یقیناً نوع بشر برای ثبات شناختی و ادراکی که می‌توان بواسطه تعمیم به آن دست یافت، ارزش قائل است. چون در نبود این امکان، جهان به سمفونی درهم و برهم و گیج کننده‌ای از نمونه‌ها و مثالهای مجزا و منفرد می‌ماند که نیازمند فراوری و تحلیل ذهنی شناختی مداوم است؛ چیزسرسام آوری که یقیناً از ظرفیت و توان شناختی محدود بشر بیرون است.

در تعریف تعمیم به عنوان یک معضل یا مسأله، همواره فرض بر این نیست که نتایج با کاربرد وسیع همواره مطلوبتر هستند (Greenwood, 1989). برای مثال، فیزیکدانی که ذرات شتاب‌دهنده را برای کشف عناصر جدید بکار می‌گیرد، انتظار ندارد که معرفی هر عنصر به دنیا مطلوب باشد. به همین ترتیب، دانشمند علوم اجتماعی ممکن است برخی اوقات تنها به دنبال نشان دادن امکان‌پذیری وجود یک اثر باشد، و اینکه متوجه شود چه مکانیزمهایی در پس اثر مشاهده‌شده وجود دارد؛ و انتظار نداشته باشد که اثر موردنظر در همه جا وجود داشته باشد. برای مثال، هنگامی‌که در جریان تغییر نگرش از طریق ارتباطات ترغیبی «اثر خفته[۶۰]» رخ می‌دهد، کاربرد این یافته آن است که تغییر با کمی‌تأخیر زمانی (و نه بلافاصله) اتفاق می‌افتد. موقعیتها و شرایطی که این اثر تحت آنها رخ می‌دهد، بسیار محدود هستند و بعید است از منظر کاربرد عمومی‌مورد توجه باشد؛ مگر برای اینکه نشان دهیم نظریه‌ای که تبیین‌کننده پدیده موردنظر است، نادرست نیست (Cook, Gruder, Henningan & Flay, 1979). آزمایشهایی که تعمیم‌پذیری محدودی دارند نیز ممکن است که به اندازه مواردی که تعمیم‌پذیری وسیعی دارند ارزشمند باشند.

با این وجود، به نظر می‌رسد نوعی تضاد میان ماهیت محدود مکان دانش علّی بدست‌آمده از هر آزمایش، و اهداف علّی تعیم یافته‌ای که تحقیق برای رسیدن به آنها انجام شده وجود دارد. کرونباخ و همکارانش (Cronbach et al., 1980; Cronbach, 1982) این بحث را به طور جدی مطرح کرده‌اند، و کارهای تحقیقاتی ایشان تا حد زیادی به تفکر موجود در خصوص تعمیم علّی یاری رسانده است. کرونباخ بر این باور است که هر آزمایش

59 Operation
60 Sleeper effect

مشتمل است بر ۱) واحدهایی[۶۱] (افرادی) که دریافت‌کننده تجربیات مورد مقایسه هستند، ۲) درمانها[۶۲] (مداخلهها)، ۳) مشاهداتی[۶۳] که بر روی واحدها انجام می‌شوند، ۴) و مختصاتی[۶۴] که مطالعه در آن انجام می‌شود. با در نظر گرفتن حرف اول هر یک از این کلمات، او یک اصطلاح ترکیبی به عنوان utos می‌سازد. utos عبارتست از افراد واقعی، درمانها، مقیاسها، و مختصاتی که در یک آزمایش به عنوان نمونه انتخاب شده‌اند. وی سپس دو مشکل همراه با تعمیم را بر می‌شمرد: ۱) تعمیم به دامنه‌ای که سؤال تحقیق در رابطه با آن پرسیده شده، چیزی که وی آن را UTOS (با حروف بزرگ) می‌نامند، و ۲) تعمیم به واحدها، مداخلهها، متغیرها، و مختصاتیست که به طور مستقیم مشاهده نشده‌اند (ص ۸۳). کرونباخ این تعمیم دوم را UTOS* می‌نامد.[۶۵]
نظریه کتاب حاضر در باب «تعمیم» -که به طور خلاصه در این قسمت به آن اشاره کرده و توضیح مفصل ان را به فصول ۱۲ و ۱۳ موکول کنیم- نظریات کرونباخ را با نظریات نگارندگان (که بر مبنای تجربیات قبلی ایشان بدست آمده است) ترکیب کرده و به نظریه‌ای بدل شده که متفاوت از هر دو ریشه نظری تشکیل‌دهنده خود است. نظریه این کتاب به دو طریق از نظریه کرونباخ تأثیر پذیرفته است. اول، نظریه این کتاب از نظر تعریف آزمایش به عنوان چیزی مشتمل بر عناصری از واحدها، مداخلهها، مشاهدات، و مختصات، پیرو نظریه کرونباخ است. اگرچه نگارندگان مکرراً به جای کلمه واحدها[۶۶] از کلمه افراد[۶۷] استفاده می‌کنند، زیرا در اغلب آزمایشهای میدانی شرکت‌کنندگان افراد هستند و نه اشیاء. همچنین با توجه به مرکزیت مشاهدات مرتبط با نتایج در هنگام آزمون روابط علّی، به جای کلمه مشاهدات از کلمه نتایج[۶۸] استفاده می‌کند. دوم، نگارندگان این موضوع را می‌پذیرند که محقق اغلب مایل به انجام دو نوع تعمیم در مورد هر یک از این پنج عنصر هستند. دو نوع تعمیمی‌که الهام گرفته از (و نه دقیقاً هم‌معنی با) دو نوع تعریف ارائه شده توسط کرونباخ است. این دو نوع تعمیم عبارتند از تعمیم‌های روایی سازه (استنباط در مورد سازه‌هایی که عملیات تحقیق نماینده آنهاست) و تعمیمهای روایی بیرونی (استنباطهایی در مورد اینکه آیا رابطه علّی برای افراد، مختصات، مداخله‌ها و متغیرهای اندازه گیری متفاوت مصداق دارد یا نه).

---

61 Units
62 Treatments
63 Observations
64 Setting

۶۵ ما با اهداف آموزشی و برای راحتی تدریس،اصطلاحات و شیوه‌های نمایشی مفاهیم استفاده شده توسط کرونباخ را بیش از اندازه ساده کرده ایم. مثلاً کرونباخ تنها از S (یعنی شکل بزرگ حرف) استفاده کرده است و نه حالت کوچه آن. بنابرایت در سیستم نمایش کرونباخ utoS وجود دارد و نه utos. او تعاریف متنوع و غیرهمگنی را از UTOSو UTOS* ارائه می‌کند. وی همچنین عبارت تعمیم دادن را با همان معنای وسیعی که در این کتاب در نظر گرفته شده، بکار نمی‌گیرد.

66 Units
67 Persons
68 Outcome

روایی سازه[۶۹]: تعمیم علّی به عنوان نماینده[۷۰]

اولین مشکل در روابط تعمیم روابط علّی برمی‌گردد به اینکه چطور می‌توان از افراد، مداخله‌ها، نتایج و مختصاتی که داده‌ها تحت آن جمع‌آوری شده، به سمت سازه‌های انتزاعی‌تر یا سطح بالاتری که این عناصر آنها را نمایندگی می‌کنند حرکت کرد. این سازه‌ها تقریباً همیشه در قالب نامها و اصطلاحاتی مطرح می‌شوند که انتزاعی‌تر از عناصریست که یک آزمایش به عنوان نمونه انتخاب شده‌اند. برچسبها (نامها) می‌تواند به هر یک از عناصر در یک آزمایش مربوط باشد؛ مثلاً، نمرات نتایج محاسبه‌شده یک تست را به عنوان دستاورد یا هوش نامگذاری می‌کند. و یا اینکه نامها می‌توانند به ماهیت رابطه میان عناصر (مشتمل بر روابط علّی) مربوط باشد. مانند زمانی که درمانهای سرطان بسته به اینکه آیا سلولهای سرطانی را مستقیماً می‌کشند و یا رشد تومور را بوسیله نامناسب کردن محیطشان متوقف می‌کنند، با نامهای کشنده[۷۱] یا مخل رشد[۷۲] نامگذاری و دسته بندی می‌شوند. آزمایشی را که در سال ۱۹۷۶ توسط فورتین و کروچ (Fortin & Kirouac, 1976) انجام شد در نظر بگیرید. مداخله یا مداخله نوعی دوره آموزشی مختصر بود که از طریق چندین پرستار ارائه می‌شد. این پرستاران تور معرفی بیمارستان محل فعالیتشان را برگزار می‌کردند، و درجریان تور حقایق و اطلاعاتی پایه‌ای درباره جراحی را در اختیار افرادی که قرار بود طی ۱۵ تا ۲۰ روز بعد از آن، در بیمارستان موردنظر جراحی شوند، قرار می‌دادند. پس از جراحی از ده مقیاس برای اندازه‌گیری نتایج استفاده شد، که مواردی از قبیل مقیاس فعالیتهای زندگی روزمره، و میزان داروی مسکن استفاده شده برای کنترل درد را در بر می‌گرفت. محقق در این مطالعه قصد بررسی این سؤال را داشت که آیا آموزش بیمار (علّت موضوعه) باعث ارتقاء فرایند بهبود و بازیابی جسمی(اثر موردنظر) در میان بیماران جراحی (جمعیت هدف) در بیمارستانها (دنیای موضوعه مختصات موردنظر) می‌شود یا نه. مثال دیگر در تحقیقات پایه‌ای اتفاق می‌افتد که در آنها این سوال مرتباً پیش می‌آید که آیا مداخله‌ها و مقیاسهای مورداستفاده در آزمایش، به‌واقع و به‌درستی نمایانگر سازه‌های علّی و اثر موردنظر نظریه هستند یا نه. یک راه برای فائق آمدن بر چالش‌های نظری و کاربردی وارده بر یک نظریه این است که حالاتی را بوجود بیاوریم که در آنها داده‌ها در واقع نماینده مفاهیم (آنگونه که مدنظر بوده است) نباشند. داده‌های تجربی غالباً محقق را وادار می‌سازند تا درک اولیه خود نسبت به ماهیت دامنه مورد مطالعه را تغییر دهد. برخی اوقات بازآفرینی مفهومی[۷۳] [یک سازه] منجر به ایجاد استنباطی محدودتر در مورد آنچه که مورد مطالعه قرار گرفته می‌شود. بنابراین، مثلاً در مطالعه فورتین و کروچ (Fortin & Kirouac, 1976) اگر جزء اطلاع رسانی درمانی به طور علّی بر بهبود پس از جراحی اثر داشته باشد، اما تور دور بیمارستان در واقع اثری نداشته

69 Construct validity
70 Representation
71 Cytotoxic
72 Cytotoxic
73 Reconceptualization

باشد، شاید لازم است علّت مورد بررسی (در اینجا آموزش بیمار) به عنوان «آموزش از طریق اطلاع‌رسانی به بیمار» بازتعریف یا بازتعیین شود. در مقابل، داده‌ها برخی اوقات می‌توانند محقق را به این سمت سوق دهند که به مواردی فکر کند که کلی‌تر از سازه‌هایی بوده که برنامه تحقیق بر اساس آن شروع شده است. بنابراین، یک تحلیلگر خلاق مطالعات آموزش بیماران ممکن است فرض کند که درمان بخشی از مداخلاتی است که از طریق افزایش «کنترل ادراک‌شده در بیمار» عمل می‌کند، و یا اینکه بهبود از جراحی می‌تواند به عنوان زیرمجموعه‌ای از «مواجهه شخصی» در نظر گرفته شود. خوانندگان مطالعه نیز می‌توانند تفسیرهای خود را از مطالعه داشته باشند، و شاید ادعا کنند که کنترل ادراک‌شده تنها یک مورد خاص از سازه‌ای کلی‌تر مانند «خودکارآمدی» است. بنابراین در طول زمان، یک تعامل یا بازی متقابل[۷۴] میان سازه‌های اولیه‌ای که محقق قصد بررسی آنها را داشته، با مطالعه‌ای که در واقع انجام شده است، نتایج مطالعه و تفسیرهای متعاقب آن رخ می‌دهد. این تعامل می‌تواند تفکر محقق نسبت به اینکه مطالعه در واقعیت و به طور اخص در سطح مفهومی، چه چیزی بدست آورده را تغییر دهد. اما هرچه بازآفرینی مفهومی‌در طول تحقیق اتفاق بیافتد، همچنان این سؤال اصلی بر جای خود باقیست که چطور می‌توان نمونه‌ای کوچک از افراد و داده‌های بدست آمده از آنها را، به سازه و مفهومی‌که این داده‌ها نمایندگی می‌کنند تعمیم داد.

روایی بیرونی: تعمیم علّی به مثابه برون‌یابی[۷۵]

دومین مسأله همراه با تعمیم استنباطها آن است که آیا یک رابطه علّی برای افراد، مختصات، مداخله‌ها، و نتایج مختلف همچنان برقرار است؟ مثل این که فردی نتایج آزمایشی که درباره اثر برنامه آموزشی در مهدکودکها بر نمرات آزمون یادگیری گرامر در میان کودکان فقیر آفریقایی-آمریکایی در شهر ممفیس در سال ۱۹۸۰ انجام شده را می‌خواند، و مایل است بداند آیا این نتایج برای ارتقاء نمرات ریاضی کودکان فقیر اسپانیایی-آمریکایی در شهر دالاس در زمان حال نیز اثربخش خواهد بود یا نه. این مثال مجدداً یادآوری می‌کند تعمیم هم‌معنی بکار بردن یک یافته در مقیاس وسیعتر نیست. در مثال بالا، تعمیم از یک شهر به شهری دیگر، یا از یک نوع خاص از مشتری به دیگر انواع است، اما چنین فرضی وجود ندارد که دالاس بزرگتر از ممفیس است، و یا اینکه کودکان اسپانیایی-آمریکایی جمعیتی وسیعتر از کودکان آفریقایی آمریکایی‌ها هستند. البته بعضی از تعمیم‌ها از کوچک به بزرگتر است. مثلاً وقتی محققی برای یک آزمایش، نمونه‌ای تصادفی از میان کل جمعیت کشور می‌گیرد، احتمالاً به دنبال تعمیم نتایج بدست آمده از نمونه به دیگر اعضاء مطالعه‌نشده جامعه است (یعنی تعمیم به کل جامعه) است. و البته از ابتدا این مهمترین منطق و دلیل برای داشتن نمونه‌ای تصادفی بوده است. به همین منوال، هنگامی‌که سیاستگذاران در مورد ادامه ارائه آموزشها و خدمات پیش‌دبستانی در سطح ملی تصمیم‌گیری

74 Interplay
75 Extrapolation

می‌کنند، آنها چندان علاقه‌ای به دانستن آنچه اختصاصاً در ممفیس اتفاق افتاده است ندارند، بلکه بدنبال این موضوع هستند که چه اتفاقی در سطح ملّی رخ خواهد داد. اما تعمیم همچین می‌تواند از جامعه بزرگتر به نمونه کوچکتر اتفاق بیافتد. کرونباخ (Cronbach, 1982) مثالهایی از یک آزمایشی ارائه می‌دهد که در آن تفاوت میان عملکرد گروههای دانش‌آموزان مدارس عمومی و خصوص مورد بررسی قرار می‌گیرد. در این مورد، مسأله موردنظر هر یک از والدین، آن است که بداند کدام نوع مدرسه برای بچه خودشان انتخاب بهتری است (نه برای کل گروه). چه از کوچک به بزرگ، و چه از بزرگ به کوچک، و چه در میان همان سطح از اندازه گیری، تمامی مثالهای مرتبط با روایی بیرونی در یک نیاز مشترک هستند، و آن اینکه می‌خواهند استنباط کنند تا چه میزان اثر موردنظر در میان طیف متفاوتی از افراد، مختصات، مداخله‌ها و یا نتایج همچنان برقرار است.

رویکردهای موجود نسبت به انجام تعمیم علّی

مسأله تعمیم دادن به هر شکلی که مطرح شود، در نگاه اول به نظر می‌رسد آزمایشها نمی‌توانند برای تعمیم چندان مفید باشند. هر آزمایش، تقریباً بلااستثناء، مجموعه محدودی از عملیات، افراد، نتایج و مختصات را در بر می‌گیرد. این سطح بالای محلی بودن عناصر تنها مختص آزمایشها نیست، بلکه مطالعات موردی، سیستم‌های پایش عملکرد و پرسشنامه‌های بازاریابی که به مجموعه‌ای تصادفی از پاسخ‌دهندگان در مراکز خرید داده می‌شود، نیز همین خصوصیات را دارند (Shadish, 1995b). حتی زمانی که پرسشنامه‌ها در میان نمونه‌هایی تصادفی که نماینده کل جمعیت کشور هستند توزیع می‌شود، نتایج تنها برای افراد همان کشور کاربرد دارد، و کمتر به جمعیت دیگر کشورها قابل تعمیم است. بعلاوه، پاسخها نیز ممکن است تحت تأثیر شرایطی که مصاحبه در آن اتفاق می‌افتد (مثلاً دم در، در اتاق نشیمن، و یا در محل کار)، زمانی از روز که جمع‌آوری داده‌ها انجام شده است، نحوه صورتبندی پرسشها، و یا نژاد، سن و جنسیت مصاحبه‌کنندگان تغییر کند. اما حقیقت آن است که این تنها آزمایشات نیستند که از نظر مسائل مرتبط با تعمیم آسیب‌پذیر هستند. بنابراین سؤالی که باقی می‌ماند آنست که این باور که آزمایشها به نحو بهتری می‌توانند میان مشخصات نمونه یک مطالعه، و استنباطهای کلی‌تر در مورد سازه و یا طیفهای متفاوت افراد، مختصات، مداخله‌ها و نتایج تناسب برقرار کند، از کجا نشات می‌گیرد؟

نمونه‌گیری و تعمیم علّی

روشی که در اکثر موارد برای دستیابی به تناسب میان مشخصات نمونه و جامعه مورد تعمیم توصیه می‌شود، استفاده از نمونه‌گیری تصادفی برای انتخاب افراد، مداخله‌ها، نتایج و یا مختصات آزمون است (Rossi, Wright, & Anderson, 1983). پیش‌فرض این است که ما به روشنی جامعه هر یک از اجزاء نمونه را می‌شناسیم، و همینطور اینکه می‌توانیم با احتمالی معین و از پیش مشخص، از میان هر کدام از این جمعیتها نمونه‌گیری کنیم. این امر

مستلزم آنست که انتخاب تصادفی، به دقت از تخصیص تصادفی تفکیک شود. انتخاب تصادفی به معنی انتخاب افراد بر اساس شانس است، بطوریکه نماینده جامعه باشند. در حالی که تخصیص تصادفی به معنای تخصیص دادن افراد به شرایط متعدد آزمایش (اعم از شرایط آزمون و کنترل) است.

در تحقیقات غیرآزمایشی علّت‌یابی نیز اغلب از نمونه‌های تصادفی استفاده می‌شود. پیمایشهای طولی[۷۶] بزرگ‌مقیاس مانند پانل مطالعات دینامیک درآمدی و یا پیمایش طولی ملّی [در آمریکا] به منظور نشان دادن جمعیت ایالات متحده (و یا بازه سنی مشخصی از آن) بکار گرفته می‌شوند. در این مطالعات ارتباط میان مقیاسهای مربوط به علّتها و اثرهای بالقوه را با وارد کردن دوره‌های زمانی و کنترلهای آماری برای گروههای غیرهم ارز[۷۷] در اندازه‌گیری‌ها بررسی می‌کنند. تمام این فرایندها به این امید انجام می‌شود که به آنچه در نمونه‌گیری تصادفی در آزمایشها بدست می‌آید، نزدیک شویم. اگرچه، مطالعاتی که در آن ابتدا از جمعیتی بزرگ نمونه‌ای تصادفی گرفته شده باشد، و سپس در درون این جمعیت تخصیص تصادفی انجام داده باشند، نسبتاً نادر است (برای مرور این مثالها نگاه کنید به فصل ۱۲). همچنین مطالعاتی که دارای انتخاب تصادفی بوده، و پس از آن یک شبه‌آزمایش با کیفیت بالا انجام شده باشد نادر است. این آزمایشها نیازمند سطح بالایی از منابع و درجه‌ای از کنترل لجیستیکی است، که غالباً انجام مطالعه را غیر اقتصادی و توجیه ناپذیر می‌کند. بنابراین بسیاری از محققین ترجیح می‌دهند تا بر مجموعه‌ای تلویحی و نانوشته از میانبرهای[۷۸] غیرآماری و شهودی در تصمیم‌گیریها تکیه کنند. میانبرهایی که امیدواریم در این کتاب بتوانیم آنها را به طور نظام‌مندتر و واضح‌تری معرفی نماییم.

احتمال انتخاب تصادفی مداخله‌ها، [گویه‌های متغیرهای] نتایج و مختصات آزمایش در یک مطالعه حتی کمتر از احتمال انتخاب تصادفی افراد است. متغیرهای نتیجه‌ای یک آزمایش را در نظر بگیرید. به چه میزان به طور تصادفی نمونه‌گیری شده‌اند؟ می‌توان تضمین کرد که مدل نمونه‌گیری دامنه در نظریه آزمون کلاسیک (Nunnally & Bernstein, 1994) فرض را بر این می‌گذارد که گویه‌های مورد استفاده برای اندازه‌گیری یک سازه، بصورت تصادفی از میان مجموعه‌ی تمامی‌گویه‌های ممکن نمونه‌گیری شده‌اند. اگرچه در واقعیت، محققین اندکی تا به حال در جریان تدوین معیارهای اندازه گیری، به طور تصادفی گویه‌ها را نمونه‌گیری کرده‌اند.

همین‌طور، این کار برای انتخاب مداخلهها و شرایط آزمون نیز انجام نمی‌شود. برای مداخله، معمولاً هیچ لیستی از مداخله‌های ممکن وجود ندارد؛ مثلاً در حوزه‌ای مانند درمان ایدز، هر روز درمانها و مداخله‌های جدیدی ابداع می‌شوند. بنابراین می‌توان گفت بطور کلی نمونه‌گیری تصادفی همواره مطلوب است، اما بندرت انجام همه جانبه آن توجیه‌پذیر است.

76 Longitudinal
77 Non-equivalent
78 Heuristics

اگرچه روشهای رسمی‌نمونه‌گیری تنها گزینه علمی‌نیستند. دو روش غیررسمی و هدفمند نمونه‌گیری نیز برخی اوقات مفید هستند – یکی نمونه‌گیری هدفمند موارد غیرهمگن، و دیگری نمونه‌گیری هدفمند موارد نوعی[۷۹]. درمورد اول، هدف در نظر گرفتن مواردی است که به طور هدفمند برای نشان دادن تنوع جنبه‌های مهم [جامعه] انتخاب می‌شوند، حتی اگر نمونه به طور رسمی‌تصادفی نباشد. در مورد دوم، هدف دادن تعریفی دقیق از انواع افراد، مداخلهها، نتایج و مختصاتی است که محقق تمایل دارد تعمیم را به آنها انجام دهد، و سپس انتخاب حداقل یک نمونه از هر طبقه که [بر اساس درک محقق] به مُد آن طبقه نزدیک باشد. اگرچه این روشهای هدفمند نمونه‌گیری نسبت به شیوه‌های رسمی‌نمونه‌گیری تصادفی عملی‌تر هستند، اما هیچ منطق آماری وجود ندارد که تعمیم‌پذیری نتایج آنها را تضمین نماید. با این حال، این روشها احتمالاً شایعترین روشهای نمونه‌گیری بکار گرفته‌شده برای تسهیل تعمیم‌پذیری بوده‌اند. یکی از مواردی که در این کتاب به آن خواهیم پرداخت توضیح همین روشها و تبیین چگونگی استفاده بیشتر از آنهاست.

هیچکدام از روشهای نمونه‌گیری به تنهایی برای حل مشکل تعمیم‌پذیری کافی نیستند. نمونه‌گیری تصادفی نیازمند تعیین دقیق جمعیتی است که نمونه از آن گرفته می‌شود، اما تعریف چنین جامعه‌ای برای برخی اهداف نمونه‌گیری مانند جمعیت مداخله‌ها دشوار است. در نمونه‌گیری هدفمند موارد غیرهمگن، انجام نمونه‌گیری برای هر یک از عناصر مختلف در یک مطالعه (افراد، مداخله‌ها، مختصات و مقیاسها) به یک میزان توجیه‌پذیر نیست. برای مثال، داشتن مقیاسهای اندازه‌گیری متنوع راحتتر و امکان‌پذیرتر از داشتن مختصات آزمون متنوع است. انجام نمونه‌گیری هدفمند از موارد نوعی نیز اغلب زمانی توجیه‌پذیر است که مُد و میانه و میانگین هدف از پیش مشخص باشد؛ اما امکان تعمیم نتایج آن نمونه به طیف وسیعتری از عناصر غیرنوعی محل سؤال و تردید خواهد بود. گذشته از آن، همانطور که کرونباخ اشاره می‌کند بسیاری از مسائل مرتبط با تعمیم‌پذیری علّی زمانی بروز می‌کنند که مدتهاست مطالعه پایان یافته. در این موارد، نمونه‌گیری تنها زمانی معنادار است که اعضای نمونه در مطالعه اولیه به اندازه کافی به طور متنوعی نمونه‌گیری شده باشند که امکان تحلیل مجدد داده‌ها وجود داشته باشد. داده‌ها باید مجدد تحلیل شوند تا بتوان بررسی کرد که آیا اثر مداخله موردنظر برای اغلب یا تمام اهدافی که امکان تعمیم نتایج به آنها زیر سوال رفته است، صدق می‌کند یا نه. اما تجمیع تعداد زیادی از منابع واریانس در قالب یک مطالعه تجربی واحد بندرت عملی و کاربردی است، و تقریباً همواره در تعارض با دیگر اهداف آزمایشی است. روشهای رسمی‌نمونه‌گیری تصادفی عموماً تنها راه‌حل‌های محدودی را برای مشکلات تعمیم‌پذیری علّی ارائه می‌کنند؛ و یک نظریه تعمیم‌پذیری استنباط علّی نیازمند ابزارهای بیشتریست.

یک نظریه بنیادی[۸۰] برای تعمیم علّی

79 Typical
80 Grounded

دانشمندان به طور روزمره در تحقیقات خود تعمیم‌های علّی انجام می‌دهند، اما تقریباً هیچگاه برای این کار از نمونه‌گیری تصادفی رسمی‌استفاده نمی‌کنند. نظریه‌ای که نگارندگان کتاب حاضر در باب تعمیم علّی ارائه می‌کنند ریشه در تجربیات واقعی ایشان در زمینه علوم دارد (Matt, Cook, & Shadish, 2000). اگرچه این نظریه در ابتدا از ایده‌هایی که ریشه در ادبیات روایی سازه و روایی بیرونی دارد نشات گرفته بود، نگارندگان پس از بررسی وسیع ادبیات موجود دریافتند که این ایده‌ها در طیف متنوعی از ادبیات مرتبط با تعمیم علمی‌عمومیت دارد (Abelson, 1995; Campbell & Fiske, 1959; Cronbach & Meehl, 1955; Davis, 1994; Locke, 1986; Medin, 1989; Messick, 1989, 1995; Rubin, 1994; Willner, 1991; Wilson, Hayward, Tunis, Bass, & Guyatt, 1995). درباره این نظریه در فصول ۱۱و ۱۳ به تفصیل بحث خواهیم کرد اما در اینجا به بیان این نکته اکتفا می‌کنیم که دانشمندان در جریان کار خود با استفاده از پنج اصل تعمیمهای علّی را ارائه می‌دهند. این پنج اصل عبارتند از:

۱) شباهت سطحی: تجانس و شباهت ظاهری میان عملیات مطالعه و مشخصات نوعی هدف مورد تعمیم ارزیابی می‌شود؛

۲) خارج کردن [81]عناصر غیرمرتبط: عناصر غیرضروری و بی‌ربط شناسایی می‌شوند، زیرا این عناصر تغییری در تعمیم‌دهی ایجاد نمی‌کنند؛

۳) تبعیض قائل شدن[82]: تبعیضهای کلیدی که موجب محدود شدن تعمیم می‌شود مشخص می‌شوند؛

۴) درون‌یابی[83] و برون‌یابی[84]: درون‌یابی بر روی مقادیری از طیف موارد نمونه‌گیری‌شده که در جریان نمونه‌گیری انتخاب نشده‌اند، انجام می‌شود. دشوارتر از آن، برون یابی‌هایی فراتر از دامنه[85] نمونه گرفته‌شده نیز کشف می‌شوند؛

۵) توضیح علّی: نظریاتی اکتشافی در مورد اثرات، علّتها و فرایندهای واسطه‌ای که برای انتقال [تعمیم] یک رابطه علّی ضروری هستند تعریف و آزمون می‌شوند.

در این کتاب نشان داده خواهد شد که چطور دانشمندان می‌توانند این پنج اصل را برای بدست آوردن نتیجه‌گیریهای تعمیم‌یافته در مورد یک رابطه علّی بکار گیرند. برخی اوقات نتیجه‌گیری در مورد سازه‌های انتزاعی‌تر (سطح بالاتر) است برای آنکه از آنها در تبیین یک رابطه بدست‌آمده در سطح نمونه استفاده کنیم. از

81 Rule out
82 Discrimination
83 Interpolation
84 Extrapolation
85 Range

این منظر، این پنج اصل مشابهات و معادلهایی در ادبیات روایی سازه (برای مثال، با محتوی سازه، با روایی همگرا و واگرا و نیاز برای وجود منطق نظری برای سازه‌ها)، در علوم شناختی و ادبیات فلسفه که به مطالعه نحوه تصمیم‌گیری افراد برای دسته‌بندی موارد در یک طبقه می‌پردازند (مثلا، تحقیقاتی که نقش مشخصات اولیه و سطحی در مقابل شباهتهای عمیقی که تعیین‌کننده عضویت در یک طبقه هستند را مطالعه می‌کنند)، دارند. اما در دیگر مواقع، نتیجه‌گیری در مورد تعمیم عبارتست از اینکه آیا یک رابطه در مقیاس کوچکتر و یا بزرگتر و برای افراد، مختصات، مداخله‌ها و نتایج مختلف همچنان معنادار است یا نه. در اینجا نیز اصلهای یاد شده دارای مشابه‌ها یا معادل‌هایی هستند، که می‌توان آنها را از نظریه‌های علمی و کاربردی بیرون کشید؛ مانند آنچه در مطالعه روابط دوزهای دارویی-پاسخ وجود دارد (شکلی از برون‌یابی و درون‌یابی) و یا جذابیتی که مکانیسم‌های توضیحی در تعمیم نتایج از حیوانات به انسان (شکلی از توضیح علّی) دارند.

دانشمندان این پنج اصل را تقریباً به طور مکرر در طول تمامی فازهای تحقیق به کار می‌گیرند. برای مثال، هنگامی‌که یک مطالعه منتشرشده را مطالعه می‌کنند و می‌اندیشند که آیا می‌توان برخی متغیرهای مطالعه موردنظر را در آزمایشگاه آنها نیز بدست آورد، در واقع در حال فکر کردن به شباهتهای مطالعه منتشر شده با مطالعه‌ای هستند که قصد انجام دادن آن را دارند. هنگامی‌که برای مطالعه‌ای جدید مفهوم‌سازی می‌کنند، در حال پیش‌بینی چگونگی طرح‌ریزی اجزاء مطالعه هستند، به نحوی که مطالعه بتواند در تناسب با اجزاء سازنده موردنظر آنها باشد. ممکن است مطالعه خود را بر پایه این فرضیات طراحی کنند که برخی واریانسها بی‌ارتباط هستند، اما دیگر واریانسها در ارتباط با تفاوتهای کلیدی نظری برای روشن کردن نحوه عملکرد مداخلهها می‌شود. در جریان تحلیل داده‌ها، محققین تمامی این فرضیات را بررسی کرده، و مختصات سازه‌های خود را برای تناسب بهتر با آنچه که داده‌های مطالعه نشان می‌دهد تعدیل می‌کنند. در قسمت مقدمه مطالعه مقالات خود تلاش می‌کنند تا خواننده را متقاعد کنند که مطالعه در ارتباط با سازه‌ای خاص است، و بخش آخر مقالاتشان به این امر اختصاص دارد که چطور نتایج بدست آمده در مورد آن سازه می‌تواند در مورد افراد، مداخلهها، نتایج و مختصات دیگر نیز صدق نماید.

بعلاوه، دانشمندان تمامی این موارد را نه فقط برای یک مطالعه، که برای تمامی مطالعات رعایت می‌کنند. آنها همواره درباره اینکه مطالعه مدنظر در مختصات ادبیات آن حوزه چه جایگاهی دارد می‌اندیشند، و به این مطلب در مقدمه مقالات خود اشاره می‌کنند. آنها تمامی پنج اصل یاد شده را در جریان مرور بر ادبیات رعایت می‌کنند. در جریان این مرور است که می‌توانند به استنباطهایی در مورد نوع رابطه علّی قابل‌دستیابی در مطالعه خود برسند.

در فصلهای ۱۱ و ۱۳ این کتاب به نظریه بنیادی تعمیم علّی بیشتر خواهیم پرداخت. تأیید این نظریه به منزله رد نمونه‌گیری تصادفی نیست و قطعاً بکارگیری این نوع نمونه‌گیری در جایی که امکانپذیر باشد، قویاً توصیه

می‌شود. اگرچه شیوه‌های نمونه‌گیری هدفمند نیز در کنار آن، و برای زمانهایی که امکان انجام نمونه‌گیری تصادفی وجود نداشته باشد، پیشنهاد می‌شود. با این همه، در این کتاب نشان خواهیم داد که نمونه‌گیری تنها یکی از روشهایی است که دانشمندان در کنار روشهایی مانند منطق کاربردی، روشهای متنوع آماری، و بکارگیری مشخصه‌های طراحی غیر از نمونه‌گیری برای انجام تعمیم‌های علّی بکار می‌گیرند.

آزمایش و علم‌پژوهی[۸۶]

مباحث فلسفی وسیعی حول نفس مسأله آزمایش و روش‌شناسی آن وجود دارد. در اینجا به طور خلاصه به برخی از این مباحث پرداخته و بعضی کاربردهای این مباحث برای روش آزمایش را ذکر می‌نماییم. اگرچه چنین به نظر می‌رسد که تمامی این مباحث فلسفی غالباً در رابطه با نحوه انجام یا عملیات آزمایش‌ها پیش می‌آیند. تاریخ آزمایش کردن به اندازه تاریخ عمر بشریت است. بنابراین بسیار پیش از تلاشهای فلسفی بشر برای درک علّیت و تعمیم علّی، نفس آزمایش وجود داشته است. حتی در طول ۴۰۰ سال گذشته نیز نوعی اثبات در مفاهیم و روشهای آزمایشی وجود داشته است. در حالی که طی همین زمان مفهوم‌سازیهای فلسفی متنوع و متفاوتی بروز و ظهوریافته، و سپس محو شده اند. همانطور که هکینگ (Hacking, 1983) عنوان می‌کند، آزمایش عمری مختص به خود دارد، و به عنوان یکی از قدرتمندترین روشهای علم برای کشف روابط علّی توصیفی بکار گرفته شده، و عملکرد بسیار رضایتبخشی از بسیاری جهات داشته؛ به طوری که جایگاه آن در علم برای همیشه تثبیت شده است. بنابراین برای توجیه عمل آزمایش، لازم نیست یک دانشمند درگیر مباحث پیچیده فلسفی در مورد عمل آزمایش شود. اما دانستن این مباحث برای شناخت بهتر آزمایش مفید است. برای مثال، وقتی در قسمتهای قبلی میان علّیت ملکولی و مولی، توصیفی و اکتشافی، یا استنباط احتمالی و قطعی تمایز قائل می‌شویم، این تمایزها می‌توانند به دانشمندان در درک بهتر آزمایش‌ها و نتایج آنها کمک نماید (Bunge, 1959; Eells, 1991; Hart & Honore, 1985; Humphreys, 1989; Mackie, 1974; Salmon, 1984, 1989; Sobel, 1993; P.A. White, 1990). در اینجا بر مجموعه وسیعتر و متفاوتی از نقدهای علمی تمرکز می‌کنیم. برخی از این نقدها به وضوح به ماهیت عمل آزمایش پرداخته، و برای یافتن نقشی تعدیل‌شده برای آن تلاش می‌کنند. این نقدها به دانشمندان برای دیدن برخی محدودیتهای آزمایش هم در علم، و هم در جامعه کمک می‌کند.

نقد کوهانی[۸۷]

کوهان (Kuhn, 1962) انقلابهای علمی را پارادایم‌هایی نامتجانس و ناکافی می‌داند که به طور ناگهانی با یکدیگر هماهنگ شده، و تبانی کردند، آن هم درست زمانی که تجمیع بطئی دانش علمی به ملغمه‌ای شتر-گاو-پلنگ شباهت داشت. هانسن (Hanson, 1958)، پلانی (Polanyi, 1959)، پوپر (Popper, 1959)، تولمین (Toulmin,

86 Metascience
87 Kuhanian Critique

1961)، فیرابند (Feyerabend, 1975)، و کویین (Quine, 1951, 1969) هرچه بیشتر به این نقد پرو بال داده و اشتباهات ناشیانه موجود در منطق اثبات‌گرایان برای ساخت فلسفه علم بر مبنای سازماندهی مجدد علم موفقی مانند فیزیک را به نمایش می‌گذارند. تمامی‌این نقدها وجود پایه‌ای مستحکم برای دانش علمی را انکار می‌کنند (بنابراین، به تبع آن، اینکه آزمایش نمی‌تواند دانش علمی مستحکمی عرضه کند). اثبات‌گرایان منطقی امیدوارند به زیربنایی دست یابند که بر پایه آن بتوانند بواسطه گره زدن نظریه صرف به مشاهدات منفصل از نظریه (با استفاده از منطق متغیرها یا محمولات[۸۸]) خلق دانش کنند. با این منطق، بسیاری از مفاهیم علمی‌که قابلیت اتصال به مشاهدات را ندارند، خارج از دایره قرار می‌گیرند. این منطق همچنین از این مسأله غافل می‌شود که تمام مشاهدات بر پایه نظریات مفهومی روش‌شناختی جمع‌آوری می‌شوند، و انجام آزمونهای مستقل از نظریه غیرممکن است.

امکان‌ناپذیربودن مشاهدات منفصل از نظریه (تز کویین-دوهم[۸۹]) به این معناست که نتایج هر آزمون (و همینطور هر آزمایش) به طور اجتناب‌ناپذیری مبهم است. نتایج را می‌توان از بسیاری جهات به نقد کشید. مثلاً اینکه فرضیات نظری مستتر در مقیاسهای اندازه‌گیری نتایج نادرست بوده است؛ و یا اینکه مطالعه فرضیات نادرستی در مورد حداقل میزان دوز لازم برای اثربخش بودن دارو داشته است. برخی از این فرضیات کوچک هستند، و براحتی می‌توان آنها را تشخیص داده و اصلاح کرد؛ مانند زمانی که یک ولتمتر اعداد اشتباهی را نشان می‌دهد، چون مثلاً مقاومت منبع ولتاژ بسیار بیشتر از متر مورد استفاده بوده است (Wilson, 1952). اما فرضیات دیگری وجود دارند که پارادایم-گونه بوده و چنان نقشی در ساختار نظریه دارند که دیگر بخش‌های نظریه بدون آنها بی‌معنا می‌شود (مثل فرضیه مرکزیت زمین در ستاره‌شناسی قبل از گالیله). از آنجایی که تعداد فرضیات موجود در هر آزمون عملی بسیار زیاد است، محققین براحتی می‌توانند فرضیاتی را بیابند که نادرست بوده و حتی می‌توانند فرضیات جدیدی طرح کنند. از این منظر، نظریات مهم و پایه‌ای بسیار کمتر از آنچه که نظریه‌پردازان آنها می‌پندارند قابل‌آزمون است. چطور می‌توان یک نظریه را آزمون کرد وقتی بجای گرانیت از سفال درست شده است.

88 Predicate logic

[توضیح مترجم: منطق محمولات، که منطق متغیرها نیز نامیده شده، بخشی از منطق صوری یا منطق نمادین است که به صورت نظام مندی رابطه‌های منطقی بین جملات را نشان می‌دهد که به سبب صرف شیوه ای رعایت می‌شود که در آن محمولات یا عبارت‌های اسمی‌بر طیفی از موضوعات به وسیلهٔ متغیرهایی همچون هر،بعضی، بدون توجه به معنا یا محتوای مفهومی‌آن‌ها یا هر محمولی به صورت خاص توزیع شده‌است. حساب محمولات متمایز از حساب جمله‌هاست که با جملات یا گزاره‌های کامل تحلیل ناشده ای با حروف عطف "اگر،آن گاه" و "یا" سروکار دارد. منطق محمولات بر خلاف منطق جمله‌ها یا منطق گزاره‌ها، با گزاره‌های اتمی‌سروکار ندارد بلکه با اجزای داخلی گزاره‌های اتمی‌سروکار دارد (نبوی، مبانی منطق و روش شناسی، ۱۳۸۹، ص۹۷)]

89 Quine-Duhem thesis

به دلایلی که بعداً به آن خواهیم پرداخت، این نقدها بیشتر در مورد مطالعه‌های منفرد مطرح است تا برنامه‌های تحقیقاتی. اما حتی در مورد برنامه‌های تحقیقاتی که مشتمل بر چندین تحقیق هستند نیز، سوگیری‌های مستمری که تشخیص داده نشوند می‌توانند منجر به شکلگیری استنباطهای نادرستی در مورد علّتها و تعمیم‌پذیری آنها شود. در نتیجه هیچ آزمایشی با قطعیت تمام نبوده، و باورها و ترجیحات فراعلمی[۹۰] همواره می‌توانند قضاوتها پایه ای موجود در باورهای جوامع علمی‌را تحت تأثیر قرار دهند.

نقدهای روانشناسی اجتماعی مدرن

جامعه‌شناسان با رویکردهای ساختارگرایی اجتماعی یا نسبیت‌گرایی معرفت شناختی، پیوسته به نقش فرایندهای فراعلمی در جریان تولید علم اشاره می‌کنند (Barnes, 1974; Bloor, 1976; Collins, 1981; Knorr-Cetina, 1981; Latour & Woolgar, 1979; Mulkay, 1979). مطالعات کاربردی آنها نشان می‌دهد که محققین غالباً نمی‌توانند به طور کامل به نرمها و قوانینی که برای [تولید] علم خوب مطرح هستند (تأثیر نپذیرفتن از احساسات و قضاوتهای شخصی[۹۱]، بی‌طرفی و به اشتراک گذاشتن اطلاعات) پایبند باشند. مباحث مطرح شده از سوی این جامعه‌شناسان همچنین به این مسأله می‌پردازد که چطور آن چیزی که به عنوان دانش علمی گزارش می‌شود، بواسطه نیروهای اجتماعی و روانشناختی و قدرتها یا فشارهای اقتصادی موجود در درون جامعه علمی و جامعه به طور کلی تعیین می‌شود (مسائلی که کمتر در تحقیقات منتشر شده به آنها اشاره می‌شود). برخی از این جامعه‌شناسان پا را فراتر گذاشته و ادعا می‌کنند که «جهان واقعی نقشی ناچیز در شکل‌دهی به دانش علمی‌دارد (Collins, 1981, p.3)».

کالینز واقع‌گرایی هستی‌شناختی[۹۲]- اینکه موجودیتهای حقیقی در جهان وجود دارند- را انکار نمی‌کند. بلکه منکر وجود واقع‌گرایی هستی‌شناختی علمی می‌شود — اینکه واقعیات بیرونی موجود می‌توانند نظریات علمی‌ما را محدود نمایند. مثلاً اگر اتم‌ها در واقع وجود دارند، آیا آنها نظریات علمی ما را تحت تأثیر قرار می‌دهند؟ اگر نظریات ما فرض بر وجود اتمها می‌گذارد، آیا واقعاً در مورد موجودیتی حقیقی، که به همان کیفیت که ما توضیح می‌دهیم وجود دارد، بحث می‌کنند؟ پاسخ رویکرد هستی‌شناختی واقع‌گرا به سؤالات بالا منفی است. کالینز بر این باور است که بیشتر اثرات پیشنهادشده در علم تحت تأثیر عوامل جامعه‌شناختی، روانشناختی، اقتصادی و سیاسی هستند. این رویکرد خارج از گروه کوچک جامعه‌شناسان چندان پذیرفته‌شده نیست، اما می‌توان به آن به عنوان موازنه‌ایی کارآمد که در مقابل این رویکرد (یا فرض ساده‌لوحانه) که مطالعات علمی واقعیت و حقیقت طبیعت را برای ما روشن می‌سازد- واقع‌گرایی ساده‌لوحانه- قرار گرفته، و آن را تا حدی تعدیل می‌کند، نگریست.

90 Extrascientific
91 Objectivity
92 Ontological realism

می‌توان گفت نتایج آزمایشها -از مفهوم‌سازی گرفته تا گزارش نتایج- هم تا حد زیادی تحت تأثیر عوامل فراعلمی است.

علم و اعتماد

اگر بخواهیم تصویری استاندارد از دانشمند را تصور کنیم، احتمالاً باید تصویر فردی بدبین را مجسم کنیم. فردی که تنها به نتایجی اعتماد می‌کند که خودش آنها را شخصاً صحت‌سنجی کرده باشد. انقلاب علمی قرن هفدهم چنین باوری را ترویج کرد که اعتماد، و علی‌الخصوص اعتماد به دولتمردان و باورهای متحجرانه- در تعارض مستقیم با علم خوب است. هر باور دیکته‌شده از سوی صاحبان قدرت و تعصبات دینی، تبدیل به مسائلی قابل پرسش و نقد شد؛ و دانشمندان نیز انجام این را وظیفه خود دانستند.

این تصویر تا حدی نادرست بود. هر مطالعه علمی به تنهایی تمرینی مبتنی بر اعتماد است. مطالعات هر کدام بر طیف گسترده‌ای از روشها، یافته‌ها، و مفاهیمی‌که طی انجام مطالعه بکار می‌گیرند، اعتماد می‌کنند. برای مثال، محقق بجای صحت‌سنجی هر یک از نظریات و روشهای آماری، آنها را به صورت پیش‌فرض مورد استفاده قرار می‌دهند. در واقع، نسبت اعتماد به بدبینی در هر تحقیق، نسبت ۹۹٪ اعتماد به ۱٪ بدبینی است. به همین ترتیب، یک محقق در طول عمر برنامه‌های تحقیق خود، بیشتر اعتماد می‌کند تا شک (Gholson et al., 1989; Shadish & Fuller, 1994). بدبینی را حتی نمی‌توان مشخصه درستی برای انقلاب‌های علمی‌گذشته نیز دانست. شاپین (Shapin, 1994) در مطالعه‌ای نشان می‌دهد که نقش «اعتماد آقامنشانه[۹۳]» در قرن هفدهم انگلستان نقشی مرکزی در شکل‌دهی به علم آزمایشی داشته است. در واقع علی‌رغم گفتمان رایج بدبینی، اعتماد در سرتاسر علم سیلان دارد.

کاربردهایی برای آزمایش

نتیجه قابل‌حصول از این نقدها، احترام و وثوق بیشتر به چندصدایی در تمامی‌زمینه‌های دانش علمی است. آزمایش را نمی‌توان پنجره‌ای شفاف که حقیقت و طبیعت را به ما می‌نمایاند در نظر گرفت. بلکه آزمایشها دانشی فرضی و دستخوش خطا در اختیار ما قرار می‌دهد. دانشی که تا حد زیادی وابسته به زمینه بوده، و از نظریات و فرضیات نظری تأثیر پذیرفته است. در نتیجه نتایج نظری تا حدی به آن نظریه‌ها و زمینه‌ها وابسته بوده، و می‌تواند با تغییر در پیش‌فرضها و زمینه‌ها دستخوش تغییر شود. از این منظر، تمامی دانشمندان از نظر هستی‌شناسی نسبیت گرا و ساختارگرا هستند؛ و تفاوتشان در میزان نسبیت‌گرایی آنهاست. نسبیت‌گراهای قوی به نظریات کالینز باور دارند، و تأیید می‌کنند که تنها عوامل فراعلمی نظریات ما را تحت تأثیر قرار می‌دهند. در

93 Gentlemanly

مقابل نسبیت‌گراهای ضعیف بر این باورند که هم جهان معرفتی[۹۴] و هم جهان ایدئولوژیک، علایق، ارزشها، امیدها و آرزو در شکل‌گیری دانش علمی ایفاء نقش می‌نمایند. بیشتر دانشمندان،که شامل نگارندگان این کتاب نیز می‌شود، احتمالاً خود را به عنوان واقعگراهای قوی (از نظر معرفت‌شناسی) و نسبیت‌گراهای ضعیف (از نظر هستی‌شناسی) تعریف می‌کنند. در واقع همان میزان از حقایق نیز که بواسطه آزمایش بر ما روشن می‌شود، از دریچه پنجره‌ای بسیار مه آلود است (Campbell, 1988).

تا همین ۳۰ سال پیش، نقش آزمایش در علم بسیار مهمتر و پررنگتر در نظر گرفته می‌شد. برای مثال، کمپبل و استنلی (Campbell & Stanley, 1963) خودشان را به عنوان وفادار به آزمایش توصیف می‌کردند: «وفاداری به آزمایش، به عنوان تنها راه‌حل برای حل و فصل مناقشات موجود درباره تجربیات آموزشی، به عنوان تنها راه صحت‌سنجی توسعه و رشد آموزشی، و به عنوان تنها راه پایه‌گذاری یک سنت تجمعی که در آن پیشرفتها و یافته‌های جدید را می‌توان بدون نگرانی نسبت به از بین رفتن دانش، و حکمت قدیمی به نفع یافته‌های جدید، معرفی کرد (ص ۲).»

هکینگ (Hacking, 1983) بر این باور است که «روشهای آزمایشی و تجربی» نام دیگر «روشهای علمی» است؛ و آزمایش بجای آنکه محلی برای نزاع و مباحثه باشد، زمینه‌ای بارور و مستعد برای مثالهایی است که روشن‌کننده مسائل پایه‌ای فلسفی هستند.

اما امروزه دیگر چنین نگاهی وجود ندارد. امروز می‌دانیم که آزمایش تلاشی بشری است (و مثل هر فعالیت بشری دیگری متأثر از تمامی نقایص و اشتباهات انسانی) با فرایند و رویه‌ای به خوبی طراحی شده که طی آن می‌توانیم به طور محدود (و نه به طور کامل)، محدودیتهایی را که تا به حال شناخته‌ایم را کنترل نماییم. برخی از محدودیتها در تمامی علوم عمومیت دارند. مثلاً دانشمندان احتمالاً بیشتر شواهد موافق با فرضیات خود را مورد توجه قرار می‌دهند، و احتمالاً کمتر شواهد متناقض با نظریات خود را در نظر می‌گیرند. آنها دچار خطاهای شناختی معمول در تصمیم‌گیری می‌شوند، و ظرفیت اندکی برای فرآوری و تحلیل مقادیر زیاد اطلاعات را دارند. آنها به فشارهای وارد آمده از سوی همکاران و همتایانشان برای همراهی با تعصبات پذیرفته‌شده و همچنین فشار ناشی از نقش‌های اجتماعی در روابطشان با دانشجویان، شرکت‌کنندگان در تحقیق، و دیگر دانشمندان واکنش نشان می‌دهند، و از آنها تاثیر می‌پذیرند. آنها تا حد زیادی بواسطه پاداشهای اقتصادی و اجتماعی محیط کار خود برانگیخته می‌شوند (متأسفانه این تا حد تقلب پیش می‌رود)، و همچنین تمام رفتارهای غیرمنطقی و نیازهای روانی انسانی را در رابطه با کار خود دارند. دیگر محدودیتها منحصراً مربوط به ماهیت خود آزمایش است. برای مثال، اگر نتایج علّی مبهم باشد، همانطور که در بسیاری از شبه‌آزمایشها به این گونه است، آزمایش‌کنندگان ممکن است علّت یا تعمیم علّی را بر پایه مشخصاتی قرار دهند که کمترین ارتباطی با منطق و

94 Ontological

روش معمول تعمیم علّی نداشته باشند. آنها ممکن است نتوانند تمامی‌گزینه‌های علّی موجّه را دنبال کنند، چون انرژی لازم را نداشته باشند و یا لازم است به موقع مطالعه‌ای را به پایان ببرند، و یا دچار این سوگیری هستند که تمایل به پذیرش شواهد همراستا با فرضیاتشان دارند. آزمایش هم یک موقعیت اجتماعی است مملو از نقشهای اجتماعی (شرکت‌کنندگان، آزمایشگرها، دستیاران ...) و انتظارات اجتماعی؛ برای مثال، اینکه افراد باید اطلاعات درستی در اختیار محقق قرار دهند. اما با این مشخصه منحصر به فرد که محقق همواره حقیقت را نمی‌گوید، و این مشکل می‌تواند مسأله ساز باشد؛ علی‌الخصص زمانی که ایماها و اشارات[۹۵] اجتماعی باعث سوءتفاهم یکی از طرفین شده، و منجر به عدم تمایل به ادامه کار و در نتیجه ترک آزمایش شود. خوشبختانه این محدودیتها اجتناب‌ناپذیر نبوده، و آموزشهای رسمی می‌توانند تا حد زیادی کمک کنند تا بر بسیاری از این مشکلات فائق آییم (Lehman, Lempert, & Nisbett, 1988). با این وجود، باید پذیرفت که رابطه میان نتایج علمی و دنیای مطالعات علمی نه ساده است و نه قابل اعتماد.

آنالیزهای روانشناختی و جامعه‌شناختی برخی از یافته‌های مرتبط به آزمایش را به عنوان اجزاء مرکزی علم در نظر می‌گیرند. می‌توان گفت آزمایش برای خودش حیاتی داشته، اما این حیات دیگر حیاتی در حاشیه امن نخواهد بود. در میان دانشمندان، باور نسبت به آزمایش، به عنوان تنها راه حل و فصل مناقشات مرتبط با علّیت، از میان رفته است. اگرچه آزمایش همچنان به عنوان روشی ارجح در بسیاری از شرایط محسوب می‌شود. به همین طریق، این باور که روشهای آزمایشی توان[۹۶] ، که غالباً در آزمایشگاهها مشاهده می‌شود به سهولت قابل‌انتقال به دنیای واقعی است، نیز از میان رفته است. در نتیجه، رویدادهای مرتبط با علم که ابعاد عمومی یافتند، نظیر حادثه چرنوبیل، مناقشات حول و حوش قطعیت آزمون DNA در آزمایشهای او.جی.سیمپسون، و شکست دهه‌ها تحقیقات گسترده‌ای که با بودجه عمومی برای یافتن درمان انواع سرطان انجام شده، درک عمومی را نیز نسبت به محدودیتهای علم افزایش داده است.

اما همچنان نباید این انتقادات را بیش از اندازه بزرگ کرد. آنها که بر علیه انجام آزمایشهای فارغ از نظریه[۹۷] ادله می‌آورند، بر این باورند که هر آزمایشی که از این روش‌ها انجام شود، به نتایج منطبق بر آرزوها و خواسته‌های محقق منتج می‌شود. البته این با تجربه محققینی که نتایج آزمایشهایشان فرساینده و ناامیدکننده بوده (از نظر تأیید نظریه دلخواهشان)، مغایرت دارد. نتایج آزمایشگاهی از خودشان چیزی نمی‌گویند، اما یقیناً در همنوایی تام با اهداف و خواسته‌های محقق نیز نخواهند بود. البته همچنان می‌توان موارد ارزشمند زیادی را در باورهای صلب و مستحکم [و پیشفرضهای ذهنی] دانشمندان دانشگاهی پیدا کرد، مواردی که عمر آنها از عمر نظریه‌های پرنوسانی که سعی در اثبات و یا تبیین آنها دارند، بیشتر است. بنابراین، بسیاری از نتایج پایه‌ای

95 Cues
96 Power experimental methods
97 Theory-free tests

در مورد جاذبه زمین، همچنان بر سیاق گذشته است، خواه در قالب چارچوب ارائه شده توسط نیوتن باشد، و خواه انیشتن. شاید بتوان گفت که حقیقت خالص وجود ندارد، اما برخی مشاهدات ارزش آن را دارند که با آنها به عنوان حقیقت رفتار کنیم.

برخی از نظریه‌پردازان علوم مانند هانسون، پلانی، کوهان و فیربند در تبیین نقش نظریه در علوم زیاده‌روی کرده‌اند، تا آنجا که شواهد و نتایج آزمایشگاهی کاملاً بلاموضوع جلوه می‌کنند. اما آزمایشهای اکتشافی‌ای که براساس نظریه‌های رسمی‌انجام نمی‌شوند، و همینطور اکتشافات غیرمنتظره آزمایشی که در حاشیه انگیزه‌های اولیه مطالعه بدست آمده‌اند، مکرراً منشاء پیشرفتهای علمی‌بزرگ بوده‌اند. آزمایشها نتایج مستحکم، قابل‌اتکاء و تکرارپذیر بسیاری فراهم آورده‌اند که بعدها موضوع نظریات مختلف را شکل داده‌اند. فیزیکدانهای آزمایشگاهی بر این باورند که داده‌های آزمایشگاهی آنها کمک می‌کند تا نظریات انتزاعی‌تر آنها قابل اعتماد باشد؛ که این خود اعتبار ویژه‌ای به آزمایشها در علم می‌دهد. البته پاره‌ای اوقات یافته‌های بعدی نشان می‌دهد که این حقایق مستحکم غیرقابل اتکاء بوده، و بیشتر مصنوع آزمایشی[۹۸] بوده اند تا حقیقت یک علّت. با این حال، این موضوع در مورد حجم بزرگی از حقایق پایه‌ای که برای مدت زمانی طولانی همچنان قابل‌اتکاء باقی مانده‌اند صدق نمی‌کند.

جهانی بدون آزمایش یا علّیت؟

با الهام از آرای مک اینتایر (MacIntyre, 1981)، فرض کنید تکه‌های علم و فلسفه از جهان پاک شوند، و ما مجبور شویم درک جدیدی از دنیا را بسازیم. به عنوان بخشی از این بازسازی، آیا مفهوم و عبارت «دلیل قابل‌دستکاری» را بازخواهیم ساخت؟ نگارندگان فکر می‌کنند که اینطور باشد. دلیل عمده آن، منافع کاربردی دلایل قابل‌دستکاری برای شکل‌دهی به یک زندگی خوشبخت و پایدار است. سوال دوم این است که آیا ما باز هم آزمایش را به عنوان روشی برای بررسی و تحقیق در مورد چنین علّتهایی اختراع خواهیم کرد؟ و باز پاسخ مثبت است. زیرا انسان همواره تمایل دارد بداند این علّتهای قابل‌دستکاری در چه شرایطی بهتر عمل می‌کنند؟ با گذشت زمان، آنها درمی‌یابند که چطور این آزمایشها را انجام دهند، و بنابراین مجدداً درگیر مسائل مرتبط با استنباطهای خلاف‌واقع، اثرات ماقبل علّت، توضیحات جایگزین، و تمامی‌دیگر عناصر و اجزاء علّیت که در این فصل به آنها پرداختیم می‌شوند. در پایان احتمالاً دوباره به چیزی شبیه آزمایش دست پیدا خواهیم کرد. کتاب حاضر قدمی‌رو به جلو در فرایند مداوم اصلاح آزمایشهاست؛ و تلاشیست در جهت بهبود دستاوردهای بدست‌آمده از آزمایشهایی که در شرایط پیچیده میدانی -چه از نظر کیفیت استنباطهای علّی حاصل از این آزمایشها، و چه از نظر توانایی ما در تعمیم این استنباطها به سازه‌ها، افراد، مختصات، مداخلهها، و نتایج مختلف- رخ می‌دهند.

98 Artifact

فصل ۲

روایی نتایج و روایی درونی

Valid: برگرفته از کلمه valide در فرانسه و از فرانسه قدیم و لاتین validus، به معنی قوی. از Valre به معنی قوی بودن. معانی مختلف این کلمه عبارتند از (۱) با ریشه و بن‌دار، برای مثال یک انتقاد ریشه‌دار؛ (۲) تولیدکننده نتایج دلخواه؛ کارآمد، مانند روشهای کارآمد؛ (۳) دارای نیروهای قانونی؛ اثربخش، مثل یک عنوان اثربخش؛ (۴) منطق: الف. مشتمل بر گزاره‌هایی که بتوان از روی آنها نتیجه‌ای منطقی بدست آورد. مثل یک بحث منطقی؛ ب. از روی گزاره‌ای به درستی استنتاج یا استنباط کردن؛ مانند نتیجه ای روا و درست۹۹

Typology: ۱. مطالعه یا تقسیم‌بندی نظام‌مند انواع یا نوعی از چیزها که مشخصاتی یا خصوصیاتی مشترک دارند. ۲. یک نظریه یا دکترین در مورد نوع ها، مانند آنچه که در مطالعات کتاب مقدس انجام می شود.

Threat : ۱. اصطلاحی برای اطلاق به نوعی تمایل به وارد کردن درد، جراحت یا تنبیه؛ ۲. شاخصی برای خطر یا صدمه؛ ۳. چیزی که ممکن است به عنوان خطر تلقی شود.

مطالعه‌ای معروف در روانشانسی قدیم در مورد اسبی به نام هانس باهوش وجود دارد که به نظر می‌رسید مسائل ریاضی را حل می‌کند و جواب را با علامت سرش نشان می‌دهد. روانشناسی به نام ایسکار فانگست عملکرد هانس را با تردید مورد بررسی قرار داد و نتیجه گرفت که او در واقع به حرکات ناملموس محقق که بر آمده از انتظارات وی بوده است، با حرکت سر پاسخ می‌داده (Pfungst, 1911). به بیان دیگر، فانگست روایی استنباط اولیه در مورد اینکه هانس باهوش مسائل ریاضی را حل می‌کرده را زیر سوال برد. در این فصل نظریه روایی که مبنای دیدگاه این کتاب در مورد تعمیم علّی را تشکیل می‌دهد، مورد بحث قرار خواهد گرفت. این فصل را با توضیح روایی هم از نظر تئوریک و هم در تجربیات علوم اجتماعی آغاز می‌کنیم. سپس به طرح یک نوع‌شناسی یا تقسیم‌بندی در

99 valid conclusion

رابطه با روایی خواهیم پرداخت. این نوع‌شناسی ایده دوگانه انواع روایی و تهدیدات به روایی را معرفی می‌نماید. فصل حاضر و فصل بعدی انواع روایی، و تهدیدات همراه با آنها را ارائه می‌نماید.

روایی

در این کتاب اصطلاح روایی، به صحت نسبی یک استنباط اشاره دارد. هنگامی که می‌گوییم چیزی رواست یا روایی دارد، به قضاوت در باب این موضوع می‌پردازیم که شواهد مرتبط تا چه اندازه استنباط موردنظر را به عنوان استنباطی صحیح و درست پشتیبانی می‌کنند. این شواهد مرتبط عموماً برآمده از یافته‌های تجربی[۱۰۰]، و همخوان با دیگر منابع دانش از جمله نظریات و یافته‌های پیشین هستند. ارزیابی روایی همواره موجد سطحی از قضاوت‌های خطاپذیر انسانیست. هیچگاه نمی‌توان مطمئن بود که تمامی استنباطهای بدست‌آمده از یک آزمایش صحیح هستند؛ و یا حتی دیگر استنباطهای ممکن به درستی و به روشی متقن ابطال[۱۰۱] شده‌اند. و به همین دلیل نمی‌توان گفت که قضاوت‌های انجام شده در خصوص روایی مطلق هستند. بلکه درجات مختلفی از روایی را می‌توان بدست آورد. بنابراین در کتاب حاضر هرگاه اصطلاح روایی بکار برده می شود، باید متوجه بود که سطحی از احتمال مدنظر است.

روایی یکی از ویژگی‌های استنباط است، و نه ویژگی طرح آزمایش یا روش. یک طرح آزمایشی واحد ممکن است در شرایط متفاوت، استنباط‌هایی با روایی بالا یا پایین به دست دهد. برای مثال، استفاده از یک طرح آزمایشی تصادفی[۱۰۲] تضمین نخواهد کرد که فرد الزاماً استنباط‌های با روایی بالا در مورد وجود یک رابطه علّی توصیفی بدست خواهد آورد. ریزشهای مختلف می‌توانند تصادفی‌سازی را مخدوش کنند، توان آماری آزمون می‌تواند کمتر از آن باشد که بتواند اثر را تایید کند، ممکن است روش آنالیز آماری نادرستی برای آنالیز داده‌ها بکار گرفته شود، و خطاهای نمونه‌گیری نیز ممکن است ما را در مورد جهت اثر به اشتباه بیاندازد. بنابراین این ادعا که آزمایش تصادفی الزاماً دارای روایی درونیست چندان صحیح نیست. به طور قطع، این نقد به تمامی دیگر روش‌های مورد استفاده در علم نیز وارد است- از مطالعه موردی گرفته تا پیمایش‌های با نمونه تصادفی. هیچ روشی نمی‌تواند تضمین‌کننده روایی یک استنباط باشد.

به همین ترتیب، از آنجا که روش‌ها تناظر یک به یک با انواع روایی ندارند، استفاده از یک روش می‌تواند به طور همزمان، بیش از یک نوع روایی را تحت تأثیر قرار دهد. شناخته‌شده‌ترین مثال در این رابطه، تصمیم برای بکارگیری آزمایش تصادفی است، که غالباً به ارتقاء روایی درونی کمک می‌کند، اما روایی بیرونی را تضعیف می‌کند. اما مثال‌های متعدد دیگری وجود دارند، مانند زمانی که تنوع بخشیدن به شرکت‌کنندگان، روایی بیرونی

100 Empirical
101 Falsified
102 Randomization

را ارتقاء می‌دهد، اما روایی نتایج آماری را کاهش می‌دهد. یا در مواردی که استاندارد کردن مداخله روایی سازه مداخله را تضمین می‌کند، اما روایی بیرونی مداخله را برای کاربرد در دیگر انواع مختصات کاربردی (که در آنها چنین استانداردهایی وجود ندارد) کاهش می‌دهد. این ماهیت تحقیقات تجربی است. به این معنا که انتخاب‌های ما در مورد طرح آزمایش، اثرات متعددی بر انواع روایی خواهد داشت، اثراتی که همواره با عواقب مورد انتظار ما مطابقت ندارد. به بیان دیگر، هر راه جدید برای حل یک مسأله می‌تواند منجر به ایجاد مسائل جدید شود. البته این منحصر به علوم نبوده و به طور کلی در مورد فعالیت‌های انسانی صدق می‌کند (Sarason, 1978).

با این حال، از دیدگاه نظریه مطرح در این کتاب، روایی به صورت مستقیم در ارتباط با مفهوم درستی و حقیقت[۱۰۳] قرار دارد. در فلسفه، سه نظریه پیشتاز در باب حقیقت وجود دارد (Schmitt, 1995): (۱) نظریه تطابق یا همترازی[۱۰۴] بر این امر دلالت دارد که یک ادعای دانشی زمانی صحیح است که مطابق با مصادیق دنیای واقعی باشد. برای مثال، این ادعا که باران می‌آید، هنگامی درست است که ما به بیرون نگاه کرده و باران را ببینیم؛ (۲) بر اساس نظریه یکپارچگی[۱۰۵]، یک ادعا هنگامی صحیح است که به مجموعه‌ای یکپارچه از ادعاها تعلق داشته باشد. برای مثال، این ادعا که استعمال ماری‌جوانا سرطانزاست هنگامی درست است که با آنچه در مورد عوارض مصرف ماری‌جوانا در سیستم‌های حیوانی مشابه بدن انسان می‌دانیم هماهنگ باشد؛ یا اینکه دیگر اشکال مصرف موادمخدر سرطانزا باشد؛ یا اینکه در میان عوامل سرطانزا موادی وجود داشته باشد که در ماری‌جوانا یافت می‌شود؛ و یا اینکه همان مکانیسم‌های فیزیولوژیکی که سیگار کشیدن را با سرطان نسبت می‌دهند، در هنگام دود کردن ماری‌جوانا فعال باشند؛ (۳) دیدگاه پراگماتیسم هنگامی یک ادعا را صحیح می‌داند، که درست دانستن [آن ادعا] به کار بیاید (استفاده داشته باشد) - برای مثال گفته می شود که «الکترونها وجود دارند». اگر استنباط چنین چیزی به مجموعه‌ای از مشاهدات معنی و مفهوم بخشیده و قدرت پیش‌بینی ما درباره چیزهایی که (در غیاب ادعای مذکور) درک آنها دشوار می‌بود را افزایش می‌دهد، این ادعا درست است. برای ایفاء نقش موردنظر ما، الکترونها لازم نیست که واقعاً وجود داشته باشند؛ بلکه صرف فرض وجود آنها نوعی ترتیب و سازماندهی ذهنی فراهم می‌آورد، و پیروی از شیوه‌ها و روالهای منضم به این ادعا در نظریه می‌تواند فواید کاربردی داشته باشد۱۰۶.

---

103 Truth

104 Correspondence

105 Coherence Theory

۱۰۶ چهارمین نظریه، که از آن به عنوان نظریه تقلیل (deflation) یاد می شود (برخی اوقات نظریه مینیمالسیتی حقیقت نیز نامیده می شود؛ Horwich, 1990)، قائل به این نیست که حقیقت الزاماً در بر گیرنده تطابق و شباهت با جهان بیرونیست و یا باید کاربردی و مفید باشد. بلکه حقیقت را نوعی ابزار کلامی سطحی و مبتذل برای « پذیرش یا همنوایی با پیشنهاداتی می داند که در قالب جملات متعدد، طولانی یا مغلق ابراز می شوند» (Schmitt, 1995, p.128). برای مثال، فرد بجای بیان مکرر موافقت خود با تمامی قوائد و اصول هندسه اقلیدوسی، می گوید «هندسه اقلیدوسی صحیح است»؛ این ادعا هیچ معنایی فراتر از آن فهرست از اصول در بر ندارد.

متأسفانه، در مورد درستی هیچکدام از این سه نظریه توافق کامل وجود نداشته، و هر سه نظریه از جنبه‌های مختلفی مورد نقد قرار گرفته‌اند. اما خوشبختانه برای اینکه بتوانیم هر کدام از این نظریات را به عنوان بخشی از یک تعریف کامل از استراتژیهای کاربردی قابل‌استفاده در ساختن سازه‌ها و تعدیل و اصلاح ادعاهای عملی تأیید کنیم، لازم نیست تنها یکی از این سه نظریه را به عنوان تنها تعریف درست انتخاب کنیم. نمود نظریه شباهت را می‌توان به وضوح در اهتمام عالم‌گیر علمی به جمع‌آوری داده به منظور ارزیابی میزان تطابق هر ادعا با حقایق دنیای واقعی، مشاهده کرد. همچنین دانشمندان عموماً در این مورد قضاوت می‌کنند که یک ادعای علمی تا چه اندازه با دیگر ادعاهای علمی در نظریه‌های موجود و یافته‌های گذشته تناسب و همخوانی دارد. و همانطور که ایزنهارت و هیو (Eisenhart & Howe, 1992) پیشنهاد می‌کنند، نتایج مطالعه‌موردی باید در تجانس و همخوانی با نظریات موجود و دانش کاربردی و نظری باشد، تا بتوان آنها (آن نتایج) را از لحاظ روایی صحیح فرض کرد. در نتیجه دانشمندان به طور سنتی به یافته‌هایی که در تناقض با بدنه موجود دانش پذیرفته‌شده باشد، با دیده تردید می‌نگرند (Cook et al. 1979). از منظر پراگماتیسم، لوتر (Latour, 1987) ادعا می‌کند دانشی می‌تواند به عنوان دانش درست در علم پذیرفته شود، که محققین بتوانند دیگر دانشمندان را به استفاده از آن مجاب نمایند؛ زیرا در جریان کاربرد مداوم است که ادعای موردنظر می‌تواند اعتبار کسب کند؛ و دستاوردهای کاربردی موردنظر حاصل شوند. این نگاه به روشنی در گفتار میشلر (Mishler, 1990) تبلور پیدا می‌کند. وی بر این باور است که «روایی روش‌های کیفی با استفاده از معیاری کارکردی ارزیابی می‌شود، که عبارتست از اینکه آیا می‌توان یافته‌های این مطالعات را مبنای مطالعات بعدی قرار داد (ص ۱۴۹)». این رویکرد را همچنین می‌توان در پاسخ متأخر به این بحث آماری-فلسفی که «برای دستیابی به علم، ارزش عملکرد بیش از همنوایی صرف با اصول فلسفی است (Casella & Schwartz, 2000, p. 427)»، مشاهده نمود.

نظریه کتاب حاضر در مورد روایی نیز از این رویکردها برای تبیین حقیقت بهره می‌برد (همانطور که تمام نظریات کاربردی روایی باید این کار را انجام دهند). نظریه نگارندگان این کتاب به ضرورت وجود همخوانی میان شواهد تجربی و استنباطهای انتزاعی پایبند بوده، و نسبت به میزان تجانس استنباط موردنظر با یافته‌های پیشین و نظریه‌های مرتبط حسّاس است. این نظریه همچنین تأکیدی پراگماتیک بر اهمیت بی‌اثر کردن[۱۰۷] تبیین‌های جایگزین دارد. تبیین‌هایی که از نظر دانشمندان حوزه‌ی خاصی از علم، می‌توانند ارزش ادعای علمی موردنظر را تضعیف نمایند- حتی اگر این تهدیدها از نظر منطقی تنها زیرمجموعه‌ای از جایگزین‌های ممکن برای ادعای علمی مورد نظر باشند. بنابراین مسیر حرکت ما به سوی حقیقت با کمک مجموعه‌ای از استراتژیها (و نه رویه‌ای واحد) ترسیم می‌شود. واقعیت آنست که نظریه شباهت با اندکی تسامح مورد استناد قرار می‌گیرد، زیرا داده‌هایی که ادعای علمی با آن مقایسه می‌شود، خود فی‌النفسه نظریه-محور بوده و بنابراین نمی‌توانند نوعی

107 Ruling out

آزمون فارغ از نظریه در رابطه با ادعای موردنظر ارائه دهند. نظریه تجانس از این نظر مورد نقد قرار می‌گیرد که لازم نیست روایت‌های متجانس الزاماً رابطه دقیقی با جهان بیرون داشته باشند (Kuhn, 1962). در نهایت، رویکرد پراگماتیسم نیز ضعف‌هایی دارد؛ چون بسیاری از باورهایی که از نظر معیارهای مختلف درست تشخیص داده می‌شوند، فایده اندکی دارند (مانند دانش ما در مورد دمای دقیق نواحی درونی یک ستاره در دوردست). بنابراین از آنجا که فلاسفه هیچکدام از این سه نظریه را به عنوان بهترین نظریه معرفی نمی‌کنند، ضرورتی ندارد دانشمندان نیز الزاماً یکی از این رویکردها را انتخاب کنند تا بتواند روایی استنباط‌های علّی و قابلیت تعمیم آنها را توجیه نمایند.

نیروهای روان‌شناختی و اجتماعی نیز تا حد زیادی آنچه که در علم به عنوان حقیقت مورد پذیرش قرار می‌گیرد (Bloor, 1997; Latour, 1987; Pinch, 1986; Shapin, 1994) را تحت تاثیر قرار می‌دهند. این در ماجرای مشهور گالیله و دادگاه‌های تفتیش عقاید بیش از هر جای دیگری نمایان است. اما پیرو نظر شاپین (Shapin, 1994) در مورد تمایز میان یک نظریه ارزیابی‌کننده[108] و یک نظریه اجتماعی در باب حقیقت، نگارندگان بر این باورند که

> *باید تعادل سست میان حقیقت، دانش و واقعیت‌های واجد اهمیت را حفظ نمود، ضمن آنکه از منافع کاربردی و مشروعیت مفهوم پردازی آزادانه تر[109] در مورد حقیقت، دفاع کرد. مفهومی که روایتی اجتماعی-تاریخی در باب حقیقت در آن مستتر باشد (Shapin, 1994, p.4).*

همانطور که بلور (Bloor, 1997) اشاره دارد، علم یک بازی برد و باخت نیست که اثرات اجتماعی و ارزشیابی-شناختی در آن جدا از یکدیگر عمل نمایند. در واقع این اثرات یکدیگر را کامل می‌کنند. نظریه‌های ارزیابی‌کننده[110] با عواملی سروکار دارند که بر آنچه *می‌باید* به عنوان حقیقت پذیرفت اثر می‌گذارند (Heider, 1944). نظریه این کتاب در باب روایی استنباطهای علّی و تعمیم‌پذیری آنها را می‌توان در زمره نظریات ارزیابی‌کننده تقسیم‌بندی نمود (Cordray, 1986). نظریه اجتماعی به کنکاش در باب عوامل خارجی اثرگذار بر آنچه ما در *واقع* به عنوان حقیقت *می‌پذیریم*، می‌پردازد. این شامل نحوه رسیدن ما به این باور که چیزی موجب ایجاد چیز دیگری شده است، نیز می‌شود. بنابراین، یک نظریه اجتماعی در مورد حقیقت ممکن است بر اساس بینش، یافته‌های روانشناسی، یا عناصر سیاسی، اقتصادی و اجتماعی محیط بنا شود. نظریه اجتماعی حقیقت در کتاب حاضر محوریت ندارد، اما در موقعیت‌های مختلف به جنبه‌هایی از آن اشاره خواهیم داشت. گرچه، حقیقت بواقع برساخته اجتماعی[111] بوده و به چیزی بیش از نظریه‌های ارزشیابی حقیقت مانند شباهت، همخوانی و

108 Evaluative
109 Liberal
110 Evaluative theories
111 Social construction

پراگماتیسم وابسته است. با این حال، نگارندگان بر این باورند که حقیقت بی‌تردید به این مسائل وابسته است، و از این رو بر این موارد تأکید می‌نمایند.

گونه شناسی روایی[۱۱۲]

دانستن کمی تاریخ می‌تواند ما را در مسیر درستی برای طراحی یک گونه‌شناسی روایی قرار دهد. کمپبل (Campbell, 1957) ابتدا روایی درونی و بیرونی را در قالب دو سؤال تعریف می‌کند: «آیا محرک آزمایشی مورد بررسی به واقع تفاوت معناداری در این نمونه خاص ایجاد کرده است؟ » (ص ۲۹۷) و «این اثر را به کدام جمعیت‌ها، مختصات، متغیرها و در نهایت چه نتایجی می‌توان تعمیم داد؟» (ص. ۲۹۷)[۱۱۳]. کمپبل و استنلی (Campbell & Stanley, 1963) نیز همین رویه را دنبال می‌کنند. به زعم آنها روایی درونی عبارتست از استنباط در مورد اینکه آیا «مداخله آزمایشی تفاوت خاصی در موضوع آزمایش ایجاد کرده است یا نه؟» (Campbell & Stanly, 1963, p.5). روایی بیرونی از سوی دیگر این پرسش را مطرح می‌سازد که «این اثر به کدام جمعیت‌ها، مختصات آزمایشی، متغیرهای مداخله، و متغیرهای اندازه گیری قابل‌تعمیم است؟» (Campbell & Stanly, 1963, p.5)[۱۱۴].

در این نوع‌شناسی، استنلی و کمپبل روایی را به چهار نوع تقسیم می‌کنند: (۱) روایی نتایج آماری، (۲) روایی درونی، (۳) روایی سازه، و (۴) روایی بیرونی. *روایی نتایج آماری*، به میزان انتخاب درست روش تحلیل آماری برای بررسی همزمان تغییرات متغیرهای مستقل و وابسته بر می‌گردد. روایی درونی به این اشاره دارد که آیا همزمانی تغییرات ناشی از رابطه علّی بوده است؟ روایی بیرونی و سازه هر دو به تعمیم ارتباط دارند. روایی سازه تعمیم از عملیات به سازه را بررسی می‌کند؛ و روایی بیرونی تعمیم از نمونه افراد، مداخله‌ها، مختصات آزمایشی، و نتایج به جامعه‌ای که سؤال تحقیق در مورد آن مطرح شده، و نیازمند تعمیم برای پاسخگویی به آن سوالات هستیم، را مورد بررسی قرار می‌هد.

---

112 A validity typology

۱۱۳به زعم کمپبل (۱۹۸۶) این تمایز تا حدی به دلیل تأکید موجود در سالهای ۱۹۵۰ بر آزمایشات تصادفی فیشری بود؛ تأکیدی که این برداشت نادرست را در دانشجویان بوجود آورده بود که تصادفی‌سازی می‌تواند تمامی تهدیدات روایی را برطرف سازد. کمپبل بر این باورست که مفهوم روایی بیرونی به این دلیل ایجاد شد که توجه محققین به آن دسته از تهدیدهایی که بواسطه تصادفی‌سازی کاهش نمی‌یابند معطوف شود؛ و بر این اساس «تهدیدات روایی درونی مواردی خواهند بود که با استفاده از تصادفی‌سازی قابل‌کنترل هستند» (p.68). اگرچه این جمله کمپبل کاملاً صحیح نیست (زیرا برای مثال، ریزش نمونه یکی از تهدیدات روایی درونی محسوب می‌شود که از طریق تصادفی‌سازی قابل کنترل نیست)، اما می‌تواند به درک مبنای فکری ایجاد این تمایز کمک نماید.

۱۱٤ روایی بیرونی برخی مواقع با روایی اکولوژیک اشتباه گرفته می‌شود. روایی اکولوژیک به طرق متفاوتی مورد استفاده قرار می‌گیرد (برای مثال، Bronfenbrenner, 1979; Burnswick, 1943, 1955). اگرچه روایی اکولوژیک در معنای اصلی خود درواقع یک نوع روایی نیست بلکه روشیست که بر تحقیقات بیشتر با نمونه هایی از مختصات آزمایشی و شرکت‌کنندگان که منعکس‌کننده اکولوژی کاربرد باشد تأکید می‌کند (البته برونفربرونر برداشتی تا حدودی متفاوت از این اصطلاح ارائه می‌دهد). این تمایز میان روایی بیرونی و درونی نیز برخی مواقع با تمایز میان مطالعه لابراتوار و میدانی اشتباه گرفته می‌شود. اگرچه تمایز دوم انگیزه‌ای برای تفکرات کمپبل بود، اما این دو تمایز منطقاً متعامد به حساب می‌آیند. اساساً، استنباط علّی بدست‌آمده از یک آزمایش میدانی روایی درونی بالایی دارد؛ و ممکن است این سؤال مطرح شود که آیا یافته‌ای که اولین بار در یک مطالعه میدانی مشاهده شده را می‌توان به آزمایشی با مختصات آزمایشگاهی تعمیم داد؟

در این کتاب، تعریف نتایج آماری و روایی درونی، همان تعریف کوک و کمپبل (Cook & Campbell, 1979) است، با این تفاوت که در تعریف ما روایی نتایج آماری بسط داده شده، و نقش اندازه اثر در آزمایشات را نیز در بر می‌گیرد. بر اساس نظر کرونباخ (Cronbach, 1982)، هر دو نوع تعمیم علّی (نمایندگی کردن یا بازنمایاندنها[۱۱۵] و برونیابی[۱۱۶]) برای تمام عناصر یک مطالعه (افراد، مداخله‌ها، نتایج و مختصات آزمایش) مصداق دارد. بر این اساس، در تعریف این کتاب، روایی سازه و روایی بیرونی به گونه‌ای تعدیل می‌شوند که در تناسب با نظر کرونباخ باشد (به جدول ۲.۱ نگاه کنید). روایی سازه عبارت است از اینکه استنباط‌های صورت گرفته در مورد افراد، مختصات، مکانیسم‌های علّت و اثر موجود در مطالعه و یا سازه‌هایی که بواسطه این نمونه‌ها نمایندگی می‌شوند، تا چه اندازه درست و قابل قبول هستند. به همین ترتیب، استنباطهای روایی بیرونی بر این امر دلالت دارد که آیا رابطه علّی مشاهده‌شده برای افراد، مختصات آزمایش، مداخله‌ها، و متغیرهای اندازه‌گیری نتایج مختلف، صدق می‌کند؟

در تعریف کوک و کمپبل (Cook & Campbell, 1979)، روایی سازه غالباً به استنباط‌های مرتبط با سازه‌های انتزاعی‌تر (سطح بالاتر) محدود می‌شد. سازه‌هایی که نمایانگر مداخله‌ها و مشاهداتی هستند که در واقع مورد مطالعه قرار گرفته‌اند. این کتاب تعریف روایی سازه را بسط داده تا افراد و مختصات آزمایش را نیز در بر بگیرد. در تعریف کوک و کمپبل (۱۹۷۹) روایی بیرونی تنها به استنباطهای مرتبط با چگونگی تعمیم یک رابطه علّی به افراد جمعیت و مختصات آزمایشی مختلف اطلاق می‌شود؛ در اینجا، این تعریف بسط داده شده، و مداخله‌ها و مشاهدات نیز در آن گنجانده شده است. به باور کوک و کمپبل(۱۹۷۹) داشتن نوعی روایی سازه مجزا برای مسائل مرتبط با علّت و اثر بصورت پراگماتیک توجیه‌پذیر است. این به دلیل توجه ویژه‌ای است که به موضوع مرکزی علّیت – یعنی کیفیت و چگونگی تبیین نظری مشخصات اثر و علّت – معطوف است. اما این تأکید بر اثر و علّت باعث شده تا پاره‌ای اوقات تعیین مشخصات جمعیت افراد و مختصات آزمایش بی‌اهمیت قلمداد شود. با توجه به نادرست بودن این رویکرد، روایی سازه باید این موضوعات را نیز در بر بگیرد. به همین ترتیب، نباید تعمیم بیرونی را به افراد و شرایط محدود کرد؛ بلکه باید این موضوع که آیا یک رابطه علّت و اثر در واریاسیون‌های مختلف علّت‌ها و اثرها همچنان معنادار است، مورد بررسی قرار بگیرد. اگرچه این تفاوتها اغلب ناچیز هستند، اما بعضی اوقات ممکن است قابل‌توجه باشند. مثال‌هایی از این نوع استنباطها در فصل سوم ارائه خواهد شد.

بحث در مورد این چهار نوع روایی که هر یک به گونه‌ای بازتعریف شده‌اند، همچنان کاربردی و پراگماتیک است، زیرا زیرا میان این چهار گونه و چهار سوال اصلی که هر محقق در مواجهه با تحقیق با آنها سروکار دارد، تطابق و

115 Representations
116 Extrapolations

شباهت وجود دارد. این چهار سوال عبارتند از: ۱) رابطه کواریانس میان علّت و اثر تا چه اندازه قوی است؟ ۲) آیا کوواریانس علّی است، یا بدون بکار بردن دستکاری نیز کوواریانس مشابهی همچنان می‌تواند وجود داشته باشد؟ ۳) مداخلهها، افراد، نتایج، و مختصات بکار گرفته شده در آزمایش با چه سازه هایی مرتبط هستند؟ رابطه علّی مشاهده شده در مقیاس کوچک تا چه اندازه قابل‌تعمیم به افراد، مداخله‌ها، مشاهدات و مختصات آزمایشی دیگر است؟ گرچه این سؤالات اغلب بسیار به یکدیگر وابسته هستند، اما لازم است تا به صورت مستقل موردبررسی قرار بگیرند؛ زیرا هر کدام مبتنی بر استنباطهایی کاملاً تفاوت قرار دارند. در پایان، با اینکه خوانندگان باید همواره به خاطر داشته باشند که «نوع‌شناسی روایی می‌تواند تا حد زیادی به طراحی آزمایش کمک می‌کند، اما جایگزین منطق و یا تحلیل انتقادی [طرح] هر مطالعه خاص نمی‌شود.» (Mark, 1986, p.63).

تهدیدهای روایی

تهدید روایی عبارتست از هر دلیلی که به واسطه آن، محقق به طور کامل یا جزئی به اشتباه افتاده باشد. این اشتباه می‌تواند هنگام استنباط درباره کوواریانس سازه‌ها، و یا در مورد اینکه آیا رابطه علّی در شرایط مختلف همچنان صدق می‌کند یا نه، روی دهد. در این فصل، خطرات و تهدیدات روایی نتایج آماری و روایی درونی را توضیح می‌دهیم. در فصل آتی نیز به بحث در مورد تهدیدات روایی بیرونی و روایی سازه خواهیم داد. تهدیدهای مطرح در رابطه با هر یک از این چهار نوع از روایی، در جریان فرایندی که همزمان، تا حدی مفهومی و تا حدی تجربیست، شناسایی شده‌اند. برای مثال، در مورد اول (مفهومی)، بسیاری از تهدیدات روایی درونی، از ماهیت استدلالها درباره استنباط‌های علّی توصیفی (که در فصل اول توضیح داده شد) نشأت می‌گیرند. در مورد دوم (تجربی)، کمپبل (Campbell, 1957) تهدیدات متعددی را از روی نقدهای وارد شده بر آزمایش‌های قبلی شناسایی نموده است. اغلب این تهدیدها به لحاظ نظری پیش‌پاافتاده به حساب می‌آیند. تهدیدهای شناسایی‌شده می‌توانند و در واقع باید در جریان تجربیات در طول زمان تغییر کنند؛ زیرا تجربیات همواره می‌توانند نیاز برای یافتن تهدیدهای جدید و بی‌اثر شدن برخی تهدیدات قدیمی را نمایان سازند. بنابراین، این کتاب یک تهدید جدید را به فهرست تهدیدکننده‌های روایی نتایج آماری می‌افزاید. این تهدید که «تخمین نادرست اندازه اثر» نامیده می‌شود، به این واقعیت اشاره دارد که در علوم اجتماعی علاوه بر انجام آزمون‌های معناداری معمول، تخمین اندازه اثر مرتبط با اثرات علّی نیز مورد تأکید است. در مقابل، اگرچه هر کدام از تهدیدهایی که توضیح خواهیم داد، در جریان انجام آزمایش‌ها اتفاق می‌افتند، اما احتمال اینکه هر کدام از آنها رخ بدهد، در شرایط و زمینه‌های مختلف متفاوت است. تهیه فهرستی از تهدیدکننده‌های روایی در واقع نوعی

میانبر[۱۱۷] است. این فهرست ثابت نبوده و ممکن است در میان تمام حیطه‌های تحقیق در علوم اجتماعی کاربرد نداشته باشد.

داشتن فهرستی از تهدیدها از آن جهت ارزشمند است، که به پژوهشگر در پیش‌بینی مشکلاتی که به طور مکرر احتمال وقوع دارد کمک کرده، و در نتیجه این امکان را فراهم می سازد تا در جریان طراحی طرح آزمایش برای خنثی کردن آنها تلاش نماید.[۱۱۸] اصلی‌ترین روشی که در این کتاب برای خنثی کردن این تهدیدها توصیه می‌شود، استفاده از کنترل‌های طرح آزمایش است، که تعداد، اثرگذاری و موجّه بودن[۱۱۹] تهدیدهایی که تا پایان مطالعه باقی می‌مانند را به حداقل می‌رساند. تمرکز اصلی این کتاب در واقع بر نحوه انجام مطالعات علّی بویژه با کمک کنترل‌های طرح آزمایشی (و نه کنترل با تعدیل‌های آماری) خواهد بود. روش دوم (یعنی کنترل بواسطه تعدیل‌های آماری) در استنباطهای علّی انجام‌شده در زمینه اقتصاد نمود بیشتری دارد. اما در مطالعات رشته آمار استفاده از این روش‌ها کمتر معمول است، و در مطالعات آماری کنترل از طریق طرح آزمایش ارجحیت دارد. تخصیص تصادفی مثالی مهم از کنترل از طریق طرح آزمون است. این کتاب عناصر طراحی آزمایش را مورد بحث قرار می‌دهد؛ عناصری که از طریق غیرموجّه کردن تفسیرهای جایگزین ادعای علّی مورد بررسی، موجب افزایش کیفیت استنباط علّی می‌شوند. در فصل ۸ نشان خواهیم داد که تخصیص تصادفی در مورد مداخله‌ها و گروه‌های کنترل چگونه و در چه زمانی می‌تواند استنباط علّی را ارتقاء دهد؛ در حالی که فصول ۴ و ۷ نشان می‌دهند که زمانی که تخصیص تصادفی امکان‌پذیر نیست، و یا با شکست مواجه شده است، چه کنترل‌هایی را می‌توان در طراحی بکار گرفت.

اگرچه، بسیاری از خطرات تهدیدکننده روایی را نمی‌توان از طریق طراحی خنثی نمود. یا به این دلیل که منطق کنترل از طریق طراحی در مورد آنها کاربرد ندارد (مانند برخی تهدیدهای روایی سازه مانند تبیین و مفهوم‌سازی[۱۲۰] ناکافی برای سازه)، و یا به دلیل آنکه محدودیت‌های عملی کنترل‌های موجود را بلا استفاده می‌نمایند. در این موارد، روش مناسب، شناسایی و کشف نقش و اثر تهدید در مطالعه است. برای انجام این کار، سه سؤال اصلی مطرح می‌شود، ۱) تهدید در این زمینه خاص چطور اثر می‌گذارد؟، ۲) آیا شواهدی وجود دارد که نشان دهد که تهدیدهای مورد نظر موجّه و جدی است؟، و ۳) آیا اثر تهدید همجهت با اثر مشاهده شده است، و بنابراین می‌تواند با اثر مورد نظر مخلوط شده و به طور اثربخش و یا کامل مشاهدات و یافته‌های مطالعه

117 Heuristics

۱۱۸ به جای اصطلاح «بی اثر کردن» یک تهدید بهتر است از اصطلاح «به حساب آوردن و در نظر داشتن» یک تهدید استفاده کنیم؛ چون نه به لحاظ نظری و نه در عمل امکان حذف کامل یک تهدید وجود ندارد. از این منظر نگارنگان با رایشارت (۲۰۰۰) همنظر هستند. وقتی از خنثی کردن تهدید صحبت به میان می آوریم اینطور برداشت می شود که حالتی وجود دارد که در آن تهدید یا وجود دارد و یا وجود ندارد؛ اگرچه در بسیاری از موارد، بحث بر سر درجات مختلف تهدید است تا وجود یا عدم وجود آن. اگرچه اصطلاح خنثی کردن چنان در ادبیات آزمایش عمومیت دارد که به دلایل شکلی و نگارشی رایج، در این کتاب همچنان از آن استفاده خواهیم نمود.

119 Plausibility

120 Explication

را تبیین کند؟ برای مثال، فرض کنید ادعا شود که تهدید گذشت زمان[۱۲۱] (اگر رویدادهایی همزمان با مداخله رخ دهند که بتوانند منجر به وقوع نتایج مشابهی شوند) تهدیدی برای روایی درونی مطالعه شبه‌آزمایشی در حال انجام، برروی اثر برنامه دولتی زنان، کودکان و نوزادان (WIC) بر ارتقاء نتایج بارداری در میان زنان با درآمد کم باشد. اولاً، لازم است بدانیم گذشت زمان چطور می‌تواند در این زمینه عمل کند. برای مثال، آیا دیگر برنامه‌های اجتماعی در دسترس هستند، و آیا زمانی که افراد تحت برنامه مذکور قرار می‌گیرند، واجد شرایط شرکت در دیگر برنامه‌ها (که می‌توانند اثر مشابه داشته باشند) نیز هستند؟ با کمی تأمل درمی‌یابیم که که برنامه‌های کمک غذایی به افراد بی‌بضاعت می‌تواند نوعی تهدید محسوب می‌شود. ثانیاً، نیاز است بدانیم آیا شواهدی وجود دارد؟ و یا حداقل انتظاری معقول- با توجه به یافته‌های قبلی یا زمینه پیشین دانش حوزه- وجود دارد که نشان دهد زنانی که واجد شرایط برنامه WIC هستند، بیشتر احتمال دارد - در مقایسه با زنانی که واجد شرایط نیستند- غذای مربوط به کمک‌های غذایی را دریافت کرده باشند؟ اگر اینطور نیست، اگرچه خطر گذشت زمان می‌تواند امکان‌پذیر باشد، اما موجّه نیست. در مورد این مطالعه خاص، بر اساس دانش زمینه‌ای موجود تهدید موجّه بود؛ چون هم برنامه WIC و هم برنامه کمک غذایی شرایط مشابهی را برای پذیرفتن افراد قرار داده بودند. سوم، اگر خطر موجّه باشد، لازم است بدانیم که آیا اثرات طرح دادن غذا به خانوارهای کم بضاعت بر نتایج بارداری، شبیه اثرات برنامه WIC است؟ اگر اینطور نیست، باز هم خطر گذشت زمان نمی‌تواند اثر مشاهده شده را توضیح دهد؛ و بنابراین، تهدیدی برای مطالعه به حساب نمی‌آید. در این مطالعه‌ی خاص، تهدید واقعی بود، چون برنامه کمک غذایی به خانوارهای کم بضاعت می‌توانست به تغذیه بهتر و متعاقباً به ارتقاء نتایج بارداری منتهی شود. در مثال‌هایی که در طول این کتاب مطرح خواهد شد، بر این سه سؤال کلیدی درباره تهدیدها تأکید خواهیم کرد.

مثال قبلی به تهدیدی اشاره داشت که پس از انجام مطالعه تشخیص داده شده. از آنجاکه تمامی محققین نمی‌توانند با سهولت کارهای تحقیقی خود را به چالش بکشند [نسبت به کار خود نگاه نقادانه داشته باشند]، این‌گونه نقدها که پس از پایان مطالعه[۱۲۲] مشخص می‌شوند، احتمالاً فراوان‌ترین منبع تهدید برای مطالعات به حساب می‌آیند. اگرچه، بهتر آن است که پژوهشگر بتواند این تهدیدها را قبل از انجام مطالعه پیش‌بینی نماید. اگر محقق بتواند تهدید را پیش‌بینی کند، اما نتواند کنترل‌های مناسبی را برای پیش‌گیری از تهدید در طرح آزمایش بگنجاند، بهترین گزینه جایگزین، محاسبه مستقیم تهدید است. به این وسیله مشخص می‌شود که آیا تهدید مذکور بواقع در مطالعه در حال انجام اثرگذار بوده است یا نه. و اگر بوده باشد، باید با استفاده از تحلیل‌های آماری بررسی کرد که آیا این تهدید می‌تواند به طور موجهی مسئول شکل‌گیری رابطه علّی و معلولی

121 History
122 post hoc

مورد نظر باشد؟ ارزیابی مستقیم اثر تهدیدات ممکن قویاً در این کتاب توصیه می‌شود (خواه بواسطه مشاهدات کمّی و خواه مشاهدات کیفی). برخی اوقات تهدیدی خاص که انتظار می‌رفت اثرگذار باشد، به شکل موردتصور عمل نکرده و یا اثر آن در خلاف جهت اثرات مشاهده شده است، و نمی‌توان اثر آن را به پای اثر مشاهده شده گذاشت (Gastwirth, Krieger, & Rosenbaum, 1994). اگرچه، باید نسبت به بکارگیری این‌گونه روش‌های مستقیم محاسبه تهدیدات در تحلیل‌های آماری، که داعیه توانایی خنثی کردن اثرات رقیب را دارند، محتاط باشیم. دلایل تکنیکی موجود برای این احتیاط در فصول بعدی این کتاب تشریح می‌شوند. اما به طور کلی، این دلایل تکنیکی به نیاز محقق برای در اختیار داشتن دانش کامل نسبت به نحوه عمل یک تهدید و نحوه محاسبه دقیق آن، مربوط می‌شوند. از آنجا که چنین دانش کاملی عموماً در دسترس نیست، نگارندگان کنترل از طریق طراحی را به کنترل از طریق آنالیز آماری ترجیح می‌دهند. اگرچه در عمل تلفیقی از هر دوی این روش‌ها مورد استفاده قرار می‌گیرد؛ و البته ترجیح آنست که بیشتر کاشیهای این تلفیق از جنس طراحی باشند. به همین منظور، کتاب حاضر عناصر متنوع طراحی آزمایش که می‌توانند در شرایط مختلف استنباط‌ها علّی، مفید واقع شده و نیاز برای تعدیلات آماری را کاهش دهند، را ارائه خواهد نمود.

در هنگام انجام تمامی این موارد، پژوهشگر باید به خاطر داشته باشد که خنثی کردن تهدیدات روایی نوعی تلاش برای ابطال[۱۲۳] است، و مشمول تمامی نقدهای ابطال که در فصل اول به آنها اشاره شد، خواهد بود. برای مثال، خنثی کردن تهدیدات موجّه در آزمایش‌ها، منوط به دانستن این خطرات است؛ و این دانش وابسته به کیفیت نظریه‌های روش‌شناختی و موضوعی مرتبط در دسترس، و همچنین مقدار دانش زمینه‌ای در اختیار پژوهشگر در خصوص موضوع تحت بررسی است. این دانش همچنین به وجود یک نظریه « توجیه پذیری[۱۲۴]» که بطور وسیع مورد پذیرش قرار گرفته باشد، بستگی دارد. به این ترتیب می‌توانیم دریابیم کدامیک از تهدیدات امکان‌پذیر متعدد در این زمینه خاص مورد مطالعه، موجّه هستند. بدون در دست داشتن چنین نظریه‌ای بسیاری از محققین به قضاوت‌های جایزالخطای خود تکیه می‌کنند (Mark, 1986; Rindskopf, 2000). همچنین باید در نظر داشت که کسب دانش نسبت به تهدیدات مرتبط با روایی به محاسبه تهدیدات از طریق شیوه‌های بدون سوگیری بستگی دارد؛ دانشی که بواسطه نظریه‌ها، آرزوها، امیدها، انتظارات و سیستمهای طبقه‌بندی ذهن آزمایشگر منحرف نشده باشد. بنابراین فرایند بی‌اثر کردن تهدیدات روایی مصداقی بارز از ابطال خطاپذیر[۱۲۵] -که در فصل اول به آن پرداخته شد- است.

روایی نتایج آماری

123 Falsification
124 Plausibility
125 Fallible falsification

روایی نتایج آماری به دو نوع استنباط آماری مرتبط، که عنصر کوواریانس در استنباط‌های علّی را تحت تاثیر قرار می‌دهند، می‌پردازد.[۱۲۶] این دو نوع استنباط عبارتند از اینکه، ۱) آیا علّت مفروض و اثر مورد نظر کواریانس دارند (همزمانی تغییرات دارند)؟، ۲) میزان (قوّت) کواریانس آنها چقدر است؟ در مورد اولین سوال، احتمال دارد به اشتباه نتیجه بگیریم که علّت و معلول دارای کوواریانس هستند، در حالیکه نیستند (خطالی نوع دوم). در مورد سوال دوم، احتمال دارد مقدار و اندازه کوواریانس را بیش از حد و یا کمتر از حد تخمین بزنیم. همچنین احتمال دارد درجه اطمینان این اندازه کوواریانس به اشتباه تخمین زده شود. با وجود اینکه تحلیل‌های کیفی کوواریانس توجیه‌پذیر و بااهمیت هستند، در این فصل نگارندگان خود را به مفهوم‌سازی کلاسیک از کوواریانس و بزرگی آن محدود می‌کنند[۱۲۷]. در ادامه با توضیحی مجمل از ماهیت آماره‌های کوواریانس شروع می‌کنیم، و سپس تهدیدات خاص موجود در مورد استنباط‌ها را مورد بحث قرار می‌دهیم.

گزارش‌دهی نتایج آزمون آماری کوواریانس

پرکاربردترین روش بکار برده شده برای بررسی کواریانس میان علّت و اثر، آزمون معناداری فرض صفر (NHST)[128] است.

به عنوان یک نمونه از این آزمون، می‌توان به پژوهشگری اشاره داشت که آزمون t را بر روی میانگین گروه‌های کنترل و آزمون، در پس‌آزمون محاسبه می‌کند، در حالی که فرض صفر معمول در این آزمون عبارتست از اینکه آیا تفاوت میان میانگین جمعیت‌هایی که نمونه‌ها از آنها گرفته شده صفر است یا نه.[۱۲۹] آزمون این فرض همراه است با عبارتی از جنس احتمال، به این مضمون که تفاوت در اندازه‌های بدست آمده، به طور شانسی رخ داده (برای مثال با احتمال p=۰/۰۳۶) در جمعیتی که در آن هیچ تفاوت میان‌گروهی وجود ندارد. به دنبال سنتی که

---

۱۲۶ در این کتاب همبستگی و کوواریانس را به جای یکدیگر استفاده می کنیم. البته اولی شکل استانداری دومی به حساب می آید. تفاوت میان این دو ممکن است برای برخی اهداف مانند زمانی که به مدل کردن فرایندهای توصیفی (explanatory processes) می پردازیم اهمیت داشته باشد.

۱۲۷ محققین کیفی اغلب بر اساس مشاهدات خود به استنباطهایی در مورد کواریانسها می رسند؛ مانند زمانی که درباره چگونگی مرتبط بودن چیزی به چیزی دیگر صحبت می کنند. باید به تهدیدات موجود نسبت به روایی اینگونه استنباطها توجه داشت. نظریه های روانشناسی درباره سوگیریهای ممکن در قضاوت نسبت به کوارریانسها و «همبستگی های وهمی و گمراه کننده» (Chapman & Chapman, 1969) تا حد زیادی می توانند به ما در درک این موضوع کمک نمایند (Crocker, 1981, Faust, 1984). اگرچه ما دانش کاملی نسبت به تمامی و یا بخش اعظمی از این عوامل تهدید کننده استنباطهای مرتبط با کوواریانس در اختیار نداریم؛ و از سوی دیگر برخی از مواردی که در گذشته شناسایی شده بودند، امروزه با نقد کشیده شده اند (Gigerenzer, 1996) زیرا به نظر می رسد این عوامل بیشتر در جریان اولین عکس العملهای فرد اثرگذار هستند. یافتن تهدیدات مترتب بر استنباطهای کیفی کاریست که بهتر است بر عهده محققین کیفی نهاده شود. با توجه به آشنایی زمینه ای آنها با این نوع تحقیقات، این محققین به نحو بهتری قادر به شناسایی و درک این تهدیدات خواهند بود.

128 Null Hypothesis Significant Test

۱۲۹ کوهن (۱۹۹۴) پیشنهاد می کند فرضیه تفاوت–صفر را فرض هیچ (nil) بنامیم تا تاکید کرده باشیم که فرض تفاوت صفر تنها فرض ممکن برای صفر کردن (nullified) نیست. این کتاب دیگر فرضهای صفر را نیز مورد بحث قرار خواهد داد. به طور سنتی، عکس فرض صفر فرض مقابل یا جایگزین خوانده می شود؛ برای مثال، اینکه تفاوت میان میانگین گروهها صفر نیست.

فیشر پایه‌گذاری کرد (Fisher, 1926, p. 504)، متأسفانه اینطور مرسوم شده که نتایج را به صورت یک دوگانه —از نظر آماری معناردار (اگر p کمتر از ۰/۰۵) و یا از نظر آماری غیرمعنادار (اگر p بزرگتر از ۰/۰۵)- تفسیر نمایند. از آنجا که غیرمعنادار به این معناست که یک علّت و اثر کوواریانس ندارند، این نتیجه‌گیری می‌تواند غلط باشد، و عواقب جدی به همراه داشته باشد. بخشی از تهدیدات روایی نتایج آماری به این برمی‌گردد که یک محقق چطور در ادعای عدم وجود اثر معنادار با استفاده از NHST به خطا افتاده است.

اگرچه مسائل ناشی از این نوع از NHST برای چندین دهه است که شناخته شده‌اند (Meehle, 1967, 1978; Rozenboom, 1960)، اما این بحث اخیراً به طور جدی‌تر مطرح شده است (Abelson, 1997; Cohen, 1994; Estes, 1997; Frick, 1996; Harlow, Muliak, & Steiger, 1997; Harris, 1997; Hunter, 1977; Nickerson, 2000; Scarr, 1997; Schmidt, 1997; Shrout, 1997; Thompson, 1993). برخی منتقدین حتی خواهان جایگزین کردن NSHT به طور کامل با دیگر گزینه‌ها هستند. البته این مباحث فراتر از موضوع این کتاب است. اما در کل می‌توان آنها را در قالب دو رشته بحث بیان کرد: ۱) دانشمندان به طور معمول NHST را به درستی متوجه نمی‌شوند، و تصور می‌کنند که p عبارتست از شانس اینکه فرض صفر صحیح باشد، و یا اینکه آزمایش باید تکرار شود (Gonzalez, Harris, & Guthrie, 1996)؛ ۲) NHST اطلاعات اندکی درباره اندازه اثر در اختیار ما قرار می‌دهد. البته برخی دانشمندان به اشتباه فکر می‌کنند معنادار نبودن آماری به معنی اثر صفر است، در حالی که در اغلب موارد، اندازه اثر چیزی متفاوت از صفر است (Lipsey & Wilson, 1993).

به همین دلیل است که بسیاری از کسانی که درباره آزمون‌های معنی‌داری آماری بحث می‌کنند ترجیح می‌دهند نتایج را در قالب اندازه اثر محدود شده در فاصله اطمینان گزارش کنند. حتی طرفداران NHST بر این باورند که این آزمون باید نقشی کمتر کلیدی در توصیف و گزارش نتایج آزمایشی داشته باشد. اما برخی (البته با تعداد کمتر) بر این باورند که NHST باید به طور کلی کنار گذاشته شود (Howard, Maxwell, & Fleming, 2000; Kirk, 1996). NHST همچنان می‌تواند برای درک نقشی که شانس می‌تواند در یافته‌های ما داشته باشد، مفید باشد (Krantz, 1999; Nickerson, 2000). بنابراین (در این کتاب) ترجیح می‌دهیم نتایج را ابتدا در قالب تخمین‌های اندازه اثر (با ۹۵٪ فاصله اطمینان)، و سپس مقدار دقیق خطای نوع اول از آزمون NHST گزارش دهیم.[۱۳۰] این کار برای هر مقایسه متمرکز میان دو شرایط آزمون و کنترل، توجیه‌پذیر به نظر می‌رسد؛ برای تمایزات مشتمل بر بیش از دو شرایط کنترل و آزمون نیز، رزنتال و روبین (۱۹۹۴) روش‌هایی را پیشنهاد می‌کنند.

---

۱۳۰ بر اساس نظر اعلام شده انجمن روانشناسان آمریکا در خصوص استنباطهای آماری «دشوار است که بتوان موقعیتی را یافت که در آن دوگانه قبول–رد بهتر از گزارش مقادیر دقیق p ، و یا حتی بهتر از آن، فاصله اطمینان باشد..... همواره در هنگام گزارش مقدار p تخمین اندازه اثر را نیز ارایه کنید. .... برای هر اندازه اثر باید تخمینهای فاصله ای ارایه شود. »(Wilkinson and the Task Force on Statistical Inference, 1999, p. 599). کوهن (۱۹۹۴) پیشنهاد می کند مطالعات منحنی های اطمینان گزارش کنند (Birnbaum, 1961) ؛ منحنی هایی که تمامی فواصل اطمینان از ۵۰٪ گرفته تا ۱۰۰٪ روی آنها خوانده شود. در این حالت دیگر نیازی نیست که یکی از این منحنی ها را انتخاب کنیم. یک برنامه کامپیوتری برای تولید این منحنی ها وجود دارد (Borerstein, Cohen, & Rothstein, in press).

اندازه اثر و ۹۵٪ فاصله اطمینان، تمامی اطلاعاتی که NHST ارایه می‌دهد را در بر می‌گیرد، اما توجه را بر بزرگی کوواریانس و دقت تخمینهای اندازه اثر متمرکز می‌کند. برای مثال، ۹۵٪ فاصله اطمینان برای ۶+-۲ ، از ۹۵٪ فاصله اطمینان برای ۶+-۵ دقت بیشتری دارد (Frick, 1996, p. 383). فاصله اطمینان همچنین به برقراری تمایز میان شرایط توان آماری پایین (در نتیجه فاصله اطمینان وسیع)، و شرایط توان آماری بالا اما اندازه اثر کوچک (شرایطی که کاربردهای متفاوتی دارد)، کمک می‌کند. گزارش آمارهای بعدی، وابستگی به تخمین‌های نقطه‌ای دقیق مورد تردید را نیز کاهش داده، و آنها را با طیف واقع‌گرایانه‌تری که به شیوه بهتری عدم‌اطمینان را بازتاب می‌دهند (حتی با وجود اینکه توضیح برای عموم را دشوار می‌نمایند) جایگزین می‌کند. بنابراین، جمله «میانگین افزایش درآمد برای هر سال ۱۰۰۰ دلار بود» با این جمله تکمیل می‌شود که «نتایج احتمالی عبارت خواهد بود از افزایش میانگین در طیفی ما بین ۴۰۰ تا ۱۶۰۰ دلار به ازای هر سال».

در تفسیر کلاسیک، سطح دقیق احتمال خطای نوع اول به ما می‌گوید نتایج مشاهده شده در آزمایش، با چه احتمالی می‌توانسته بواسطه شانس- در جامعه‌ای که فرض صفر در آن صحیح است- رخ داده باشد (Cohen, 1994). از این نظر،NHST اطلاعاتی در این خصوص که نتایج می‌توانسته کاملاً شانسی بدست آمده باشد، در اختیار ما قرار دهد (شاید این جالبترین فرضیه نباشد اما فرضیه‌ای است که به طور معمول به خواننده در مورد آن اطلاعات می‌دهیم). تفسیر جالبتر می‌تواند این باشد که سطح احتمال به ما می‌گوید با چه میزان اطمینان می‌توانیم در مورد سه ادعا قضاوت کنیم؛ این سه ادعا عبارتند از ۱) علامت (یا جهت) اثر در جمعیت مثبت است (مداخله A بهتر از مداخله B عمل کرده)؛ ۲) علامت (جهت) اثر منفی است (مداخله B بهتر از مداخله A عمل کرده)؛ ۳) علامت نامعلوم است. هرچه مقدار P کمتر باشد، احتمال کمتری وجود دارد که نتایج ما درباره جهت و علامت اثر در جمعیت اشتباه باشد؛ و اگر p≥ ۰/۰۵ باشد (یا معادل آن، اگر بازه اطمینان شامل صفر باشد)، آنگاه نتایج ما در خصوص اثر علامت، به سختی قابل پیش‌بینی است.

در هر حال، تفسیر مقدار p هر چه باشد، تمامی این موارد ما را از این نتیجه‌گیری سهل‌انگارانه که «اثر وجود دارد» یا «اثر وجود ندارد» باز می‌دارد. نگارندگان این کتاب بر این باورند که NHST به تدریج نقش کوچکتری در علوم اجتماعی ایفاء خواهد کرد، اگرچه هیچ رویکرد جدیدی نیز کامل و بی نقص نیست.[۱۳۱] همانطور که ابلسون اخیراً اظهار کرده:

---

۱۳۱ به عنوان جایگزینی دقیقتر و مکملی برای NHST و گزارش اندازه اثر با فاصله اطمینان، می توان از آماره های بیزین استفاده کرد. به جای پذیرش یا رد ساده فرض صفر، رویکردهای بیزین عموماً نتایج حاصل از یک مطالعه را برای به روز رسانی دانش موجود به کار می گیرند. این کار خواه از طریق مشخص کردن انتظارات در مورد نتایج مطالعه قبل از شروع مطالعه، و خواه به طور گذشته نگر از طریق اضافه کردن نتایج از یک آزمایش به مجموعه ای موجود از آزمایشات که با استفاده از روشهای بیضی نتایج را به روز آوری کرده اند، انجام می شود. روش دوم بسیار به رویه های متاآنالیز اثر تصادفی (که در فصل ۱۳ به آن پرداخته ایم) نزدیک است. تا همین اواخر، آمار بیضی با اکراه و امساک بکار گرفته می‌شد. بخشی به دلیل آنکه روشهای بیضی از نظر

«هر اتفاقی برای NHST بیافتد؛ بگذارید دست از جلوه دادن تحلیل آماری به عنوان یک فرایند کامل و بدون خطا پایان دهیم. ما در دریایی از عدم اطمینان مستغرق‌ایم که با سیل خطاهای نمونه‌گیری و اندازه‌گیری موّاج است، و هیچ رویه عینی که مستقل از سوگیریها بوده، و با آن بتوان از قضاوت‌های انسانی دوری کرد، و تضمین‌کننده تفسیر درست نتایج باشد، در اختیار نداریم (Abelson, 1997, p. 13)»

جدول ۲.۲: تهدیدهای روایی نتایج آماری: دلایلی که بواسطه آنها استنباط در مورد کوواریانس میان دو متغیر می‌تواند ناصحیح باشد

۱. *توان آماری اندک:* آزمایشی با توان آماری ناکافی می‌تواند به غلط به این نتیجه‌گیری منتهی شود که رابطه میان مداخله و نتایج معنادار نیست.
۲. *پیش‌فرض‌های نقض‌شده آزمون‌های آماری:* بی‌توجهی و نقض پیش‌فرض‌های آزمون آماری می‌تواند به تخمین بیش از اندازه یا کمتر از اندازه، و معناداری یک اثر بیانجامد.
۳. *دستچین کردن*[۱۳۲] *و مسأله نرخ خطا:* آزمون‌های مکرر برای روابط معنادار، اگر با اصلاح برای تعداد آزمون انجام نشود، می‌تواند به طور تصادفی موجب افزایش معناداری آماری شود.
۴. *قابل اعتماد نبودن مقیاس‌ها:* خطاهای اندازه‌گیری رابطه میان دو متغیر را تضعیف می‌کنند، و رابطه میان سه یا بیشتر متغیر را تقویت و تضعیف می‌نمایند.
۵. *محدودیت دامنه*[۱۳۳]*:* دامنه محدود یک متغیر، معمولاً رابطه میان آن متغیر و دیگر متغیرها را تضعیف می‌نماید.
۶. پایایی پایین شیوه اجرای مداخلهها: اگر یک مداخله که باید به عنوان روشی استاندارد بکار گرفته شود، به صورت بخشی، و تنها برای بعضی پاسخ‌دهندگان استفاده شود، اثرات ممکن است در مقایسه با زمانی که مداخله به صورت کامل به کار گرفته می‌شد، کمتر از حد تخمین زده شوند.
۷. واریانس خارجی[۱۳۴] (غیرمرتبط) در مختصات آزمایشی: برخی مؤلفه‌های یک شرایط آزمایشی موجب افزایش خطا می‌شوند، و تشخیص وجود اثر را دشوار می‌نمایند.

محاسباتی سنگین هستند و برنامه های قابل قبول اندکی برای انجام آنها وجود داشت. انتظار می رود شاهد استفاده گسترده‌تر و با فراوانی بیشتر از آمار بیضی در دهه‌های آینده باشیم.

132 Fishing

133 Range

134 Extraneous

۸. عدم تجانس و ناهمگونی افراد: افزایش نوسان در متغیر نتیجه در درون هر یک از شرایط (کنترل یا آزمون) واریانس خطا را افزایش می‌دهد، و در نتیجه تشخیص یک رابطه را دچار مشکل می‌سازد.

۹. *تخمین نادرست اندازه اثر:* برخی آماره‌ها به طور نظام‌مند اندازه یک اثر را بیش از حد یا کمتر از حد تخمین می‌زنند.

تهدیدات روایی نتایج آماری

جدول ۲-۲ فهرست تهدیدهایی را نشان می‌دهد که روایی نتایج آماری را به خطر می‌اندازند. به بیان دیگر، این تهدیدات عللی هستند که موجب می‌شوند محققان در دستیابی به استنباط‌های درست در مورد وجود و اندازه کوواریانس میان دو متغیر دچار خطا شوند.

توان آماری پایین

توان آماری عبارتست از توانایی یک آزمون برای تشخیص رابطه‌ای که، در واقع، در جامعه مورد بررسی وجود دارد. به طور قراردادی، توان آزمون به صورت احتمال آن که یک آزمون آماری فرض صفر را رد کند، تعریف می‌شود، هنگامی که در حقیقت این فرض نادرست بوده باشد (Cohen, 1988; Lipsey, 1990; Maxwell & Delaney, 1990). هنگامی که یک مطالعه توان آماری کمتری دارد، تخمین‌های اندازه اثر دقت کمتری دارند (فاصله اطمینان وسیعتری دارند)، و NHST ممکن است به اشتباه به این نتیجه منتهی شود که علّت و اثر کوواریانس ندارند. برنامه های کامپیوتری به سادگی می‌توانند توان آماری را محاسبه کنند، البته اگر بتوانیم اندازه نمونه، خطای نوع اول و نوع دوم و اندازه اثر را بدرستی تخمین بزنیم (Borenstein & Cohen, 1988; Dennis, Lennox, & Foss, 1997; Hintze, 1996; Thomas & Krebs, 1997). در علوم اجتماعی مقدار خطای نوع اول معمولاً برابر ۰/۰۵ در نظر گرفته می‌شود، اگرچه برخی مواقع دلایل خوبی وجود دارند برای اینکه بتوان از این حد عدول کرد. برای مثال، هنگامی که یک داروی جدید را از نظر عوارض جانبی مخرب آن آزمون می‌کنیم، یک سطح خطای نوع اول محتاتانه‌تر مناسبتر است (مثلاً ۰/۰۲). همینطور معمول است که خطای نوع دوم را در سطح ۰/۲ در نظر می‌گیرند، و بنابراین توان آزمون برابر با ۰/۸ (۱-ß) خواهد بود. اندازه اثر هدف از روی اثری که به لحاظ نظری معنادار یا از نظر کاربردی با اهیمت است، استنباط می شود (Cohen, 1996; Lipsey, 1990). انحراف معیار مورد نیاز برای محاسبه اندازه اثر معمولاً از روی مطالعات پیشین یا مطالعات پایلوت استخراج می‌شود. اگر توان آماره برای یافتن اثر با اندازه مشخص خیلی پایین باشد، قدمهایی می‌توان برای افزایش توان برداشت.

با توجه به اهمیت کلیدی توان در طراحی کاربردی آزمایش، جدول ۳-۲ برخی از عوامل اثرگذار بر توان آماری را به طور خلاصه به نمایش می‌گذارد. این موارد از نظر قابل استفاده بودن، توجیه‌پذیر بودن، انتظار از کاربرد آنها، و معایبشان مورد بررسی قرار می‌گیرند.

جدول ۳-۲: روش‌های افزایش توان آماری

روش	پیشنهادات
استفاده از جفت کردن (همتا کردن)[۱۳۵]، طبقه بندی‌کردن[۱۳۶] و بلوک‌بندی کردن[۱۳۷]	۱. اطمینان حاصل کنید که متغیرهای مورداستفاده برای همتا کردن، طبقه‌بندی کردن یا بلوک‌بندی کردن با متغیرهای نتیجه‌ای همبستگی دارند، و یا متغیری را بکار بگیرید که زیرتحلیلها[۱۳۸] بر اساس آن برنامه‌ریزی شده است (Maxwell, 1993) ۲. در صورتی که تعداد افراد کم باشد، استفاده از همتا کردن باعث کاهش توان می‌شود (Gail et al., 1996)
متغیرهای تصادفی کمکی[۱۳۹] را محاسبه و اصلاح کنید	۱. متغیرهای تصادفی کمکی همبسته با نتایج را محاسبه کنید، و آنها را برای تحلیل آماری تعدیل و کنترل کنید (Maxwell, 1993) ۲. در هنگام افزودن بر متغیرهای تصادفی کمکی و افزایش اندازه نمونه، به میزان افزایش هزینه برای بدست‌آوردن توان آماری بیشتر توجه کرده، و میان هزینه اضافی و توان بیشتر توزان برقرار نمایید (Alison, 1995; Allison et al., 1997) ۳. متغیرهای تصادفی کمکی‌ای را انتخاب کنید که در کنار دیگر متغیرهای تصادفی کمکی زائد و غیرضروری نباشند (McClelland, 2000) ۴. متغیر تصادفی کمکی بکار گرفته شده برای آنالیز متغیرها را برای بلوک‌بندی، همتا کردن یا طبقه‌بندی کردن مورد استفاده قرار دهید
از نمونه با اندازه بزرگتر استفاده کنید	۱. اگر تعداد افراد شرایط مداخلهها ثابت باشد، تعداد شرکت‌کنندگان گروه کنترل را افزایش دهید ۲. اگر بودجه ثابت، و مداخله از کنترل گرانتر است، توزیع بهینه منابع برای بدست آوردن توان را محاسبه کنید (Orr, 1999)

135 Matching
136 Stratifying
137 Blocking
138 Subanalysis
139 Covariate

	۳. با اندازه نمونه کل ثابت، هرگاه مجموعه‌ها[۱۴۰] به هر یک از شرایط کنترل و آزمون تخصیص داده می‌شوند، تعداد مجموعه‌ها را افزایش، و تعداد افراد داخل مجموعه‌ها را کاهش دهید
استفاده از تعداد نمونه مساوی برای هر خانه (در شرایط کنترل و آزمون)	۱. تقسیم نامتوازن سلول کمتر احتمال دارد توان آماری را تحت تأثیر قرار دهد، مگر اینکه نسبت تقسیم به بیشتر از ۲:۱ برسد (Pocock, 1983) ۲. برای برخی اثرات، تقسیم‌های نامساوی اندازه نمونه می‌تواند پرتوانتر باشد (McClelland, 1997)
بهبود مقیاس‌های اندازه‌گیری	۱. پایایی مقیاسها را افزایش دهید، و یا از مدل‌سازی مکنون متغیرها بهره بگیرید ۲. محدودیت بی‌دلیل دامنه[۱۴۱] را از میان ببرید (برای مثال، به ندرت مقیاس‌های پیوسته را به مقیاس‌های دوتایی تبدیل کنید) ۳. بیشتر منابع خود را روی پس‌آزمون متمرکز کنید تا پیش‌آزمون (Maxwell, 1994) ۴. موج های اضافی اندازه‌گیری اضافه کنید (Maxwell, 1998) ۵. از اثرات سقف و کف اجتناب کنید
قدرت مداخله را افزایش دهید	۱. تفاوت دوز (مقدار) مداخله میان شرایط آزمون و کنترل را افزایش دهید ۲. انتشار و جابجایی[۱۴۲] میان شرایط کنترل و آزمون را کاهش دهید ۳. توزیع مطمئن[۱۴۳]، دریافت، و پایبندی به استفاده از مداخله را تضمین کنید
تغییر-پذیری[۱۴۴] مداخله را افزایش دهید	۱. دامنه سطوح مداخله در حال آزمون را افزایش دهید (MacClelland, 2000) ۲. در پاره‌ای مواقع، از حدهای انتهایی[۱۴۵] مداخله تعداد نمونه بیشتری بگیرید (MacClelland, 1997)
از طراحی درون-گروهی[۱۴۶] استفاده کنید	۱. این طرح‌ها خارج از آزمایشگاه کمتر شدنی و توجیه پذیر هستند ۲. خطراتی مانند خستگی، تمرین، تجربه و آلودگی تهدیدی برای این طرح‌ها محسوب می شوند
از شرکت‌کنندگان متجانس و مشابه برای	۱. می‌توانید در مورد تعمیم‌پذیری مسامحه کنید

140 Aggregates
141 Range
142 diffusion
143 Reliable
144 Variability
145 Extreme
146 Within-subject

پاسخگویی به مداخله‌ها استفاده کنید	
عناصر نامربوط مختصات تصادفی را کاهش دهید	۱. می‌توانید در مورد برخی انواع تعمیم‌پذیریها مسامحه کنید
اطمینان حاصل کنید که آزمون‌های آماری قدرتمندی بکار گرفته‌شده، و شرایط و پیشفرض‌های هر یک از آنها رعایت شده است	۱. عدم پایبندی به پیش‌فرض‌های آزمون برخی اوقات توان را افزایش می‌دهد (مثلاً اینکه با افراد مرتبط مانند افراد مستقل رفتار کنیم)، بنابراین شما باید رابطه میان پیش‌فرضها و توان را بدانید ۲. تغییرشکل[۱۴۷] داده‌ها برای نرمال‌سازی داده‌های می‌تواند توان را افزایش دهد، اگرچه نمی‌تواند اثر چندانی بر خطای نوع اول داشته باشد (McClelland, 2000) ۳. روش‌های آماری جایگزین را نیز در نظر بگیرید (برای مثال، Wilcox, 1996)

اگر بخواهیم بر اساس ادبیات موجود قضاوت کنیم، فراوانی آزمایشهای با توان آماری پایین در میان مطالعات موجود قابل‌توجه است. برای مثال، کازدین و باس (Kazdin & Bass, 1989) دریافتند که در اغلب مطالعات روان‌درمانی که در آنها دو مداخله مقایسه شده‌اند، توان آماری بسیار پایین بوده است (همچنین نگاه کنید به Freiman, Chalmers, Smith, & Kuebler, 1978; Lipsey, 1990; Sedlmeier & Gigerenzer, 1989). بنابراین توان آماری پایین منشاء اصلی نتایج اشتباه در مورد فرض صفر در مطالعات منفرد است. اما هنگامی که اثرات کوچک هستند، غالباً افزایش توان (به میزان کافی) با استفاده از روش‌هایی که در جدول در جدول ۳-۲ مورد اشاره قرار گرفت، امکان پذیر نیست. به همین دلیل است که امروزه ترکیب و سنتز مطالعات متعدد (به فصل ۱۳ نگاه کنید) به عنوان راهی به سوی آزمون‌های قویتر اثرات کوچکتر توصیه می‌شود.

پیش‌فرض‌های نقض‌شده آزمون‌های آماری

اگر پیش‌فرض‌های آزمون‌های آماری نقض شوند، استنباطها و نتایج مربوط به کوواریانسها ممکن است نادرست از آب دربیاید. نقض برخی پیش‌فرض‌ها می‌تواند کمتر آسیب‌زا باشد. برای مثال، یک آزمون t دوطرفه نسبت به نقض شرط نرمال بودن توزیع‌ها، آسیب‌پذیری کمتری دارد، البته اگر اندازه نمونه‌ها بزرگ بوده، و نمونه‌ها برابر باشند، و تنها خطای نوع اول مسأله‌ساز باشد (Judd, McClelland, & Culhane, 1995 برای خطای نوع دوم نگاه

147 Transformation

کنید به، Wilcox, 1995). اما نقض برخی دیگر از پیش‌فرض‌ها جدی‌تر است. برای مثال، در صورتی که مشاهدات مستقل نباشند، نتایج و استنباط‌های مرتبط با کوواریانس می‌تواند دچار مشکل شود. مثلاً، کودکان حاضر در یک کلاس واحد بیشتر احتمال دارد با یکدیگر در ارتباط باشند تا کودکانی که به طور تصادفی انتخاب می‌شوند؛ بیماران مراجعه‌کننده به یک پزشک، و یا همکارانی که در یک مکان مشترک کار می‌کنند بیشتر احتمال دارد به یکدیگر شبیه باشند. این عامل تهدیدزا اغلب وجود دارد، و پیش‌فرض استقلال را نقض می‌نماید. این مسأله می‌تواند سوگیری‌های قابل‌توجهی در تخمین خطای استاندارد –که اثر خالص و دقیق آن بسته به طراحی و نوع وابستگی متفاوت است- ایجاد نماید (Judd et al., 1995). در بسیاری موارد که افراد درون مجموعه لانه گزینی[۱۴۸] کرده‌اند (برای مثال کودکان در برخی مدارس یک مداخله را می‌گیرند، و دیگر کودکان در مدرسه‌ای دیگر گروه کنترل را تشکیل می‌دهند)، این سوگیری پیش می‌آید؛ خطای نوع اول به شدت افزایش می‌یابد و محقق چنین نتیجه‌گیری می‌کند که تفاوت «معناداری» از اعمال مداخله حاصل شده است. خوشبختانه در سال‌های اخیر پیشرفت‌هایی در زمینه اقدامات اصلاحی آماری و برنامه‌های کامپیوتری برای حل این مشکلات حاصل شده است (Bryk & Raudenbush, 1992; Bryk, Rauenbush, & Congdon, 1996; DeLeeuw & Kreft, 1986; Goldstein, 1987).

دستچین کردن و مشکل نرخ‌ها

در صورتی‌که استنباط‌های کوواریانس در نتیجه دستچین کردن از درون داده‌ها برای یافتن اثری معنادار در آزمون فرضیه صفر، و یا برای دنبال کردن مسیرهایی که در خود داده ها پیدا شده انجام شوند، این استنباط‌ها می‌تواند واجد خطا و اشکال باشد. احتمال بروز این خطا زمانی که محققین متعددی یک داده واحد را تحلیل می‌کنند افزایش می‌یابد (Denton, 1985). هنگامی که خطای نوع اول برای یک آزمون برابر با ۰/۰۵ باشد، نرخ خطا برای مجموعه‌ای از همان آزمون کاملاً متفاوت است؛ و با افزایش تعداد آزمون، احتمال خطای نوع اول افزایش می‌یابد. اگر سه آزمون با آلفای ۰/۰۵ انجام شود، آنگاه آلفای واقعی (یا احتمال ارتکاب خطای نوع اول برای مجموع سه آزمون) چیزی حدود ۰/۱۴۳ خواهد بود؛ با ۲۰ آزمون ۰/۶۴۲ و با پنجاه آزمون ۰/۹۲۳ (Maxwell & Delaney, 1990). در نتیجه، خصوصاً در مواردیکه تنها یک زیرمجموعه از نتایج گزارش شود، نتایج تحقیق می‌تواند گمراه‌کننده باشد.

آسانترین رویه اصلاحی استفاده از روش محافظه‌کارانه اصلاح بانفرونی است، که در آن نرخ خطای نوع اول کلی برای یک مجموعه آزمون (a=۰/۰۵) را بر تعداد آزمون موجود در مجموعه تقسیم می‌کند، و آلفای اصلاح‌شده بانفرونی را برای هر کدام از آزمون‌های مجموعه مورد استفاده قرار می‌دهد. این رویه تضمین‌کننده آن خواهد بود

148 Nested

که نرخ خطای کلی تمام آزمونها از نرخ ۰/۰۵ تجاوز نمی‌کند. از دیگر روش‌های اصلاحی، می‌توان به استفاده از آزمون‌های محافظه‌کارانه متعدد پیگیری[۱۴۹] در تحلیل واریانس[۱۵۰] و یا بکارگیری تحلیل واریانس چندمتغیره[۱۵۱] (اگر متغیرهای وابسته متعدد در حال آزمون‌شدن هستند) اشاره داشت (Maxwell & Delaney, 1990). برخی نقدهای مرتبط با آزمون فرض صفر این اصلاحات را نفی می‌کنند؛ و چنین استدلال می‌کنند که به طور معمول تمایل به نادیده‌گرفتن اثرات کوچک وجود دارد، و اصلاحات محافظه‌کارانه این احتمال را افزایش می‌دهد.
گزارش کردن اندازه اثر، فاصله اطمینان و مقدار دقیق p توجه و تأکید را از تصمیم در مورد «معناداری-غیر معناداری فرضیات»، به سوی اطمینان در مورد اندازه و جهت احتمالی اثر سوق می‌دهد. نقدهای دیگر به این موضوع اشاره دارند که اگر نتایج برای تمام آزمون‌های آماری به طور کامل گزارش شود، خوانندگان خودشان می‌توانند شانس یا احتمال نتایج معنادار را ارزیابی کنند (Greenwald et al., 1996). اما معمولاًبه دلیل فضای محدود در مقالات، و همچنین تمایل نویسندگان به محدود کردن گزارش نتایج به مواردی که بیانگر داستانی جالب است، گزارش کامل نتایج ارائه نمی‌شود. بنابراین در اغلب موارد دستچین کردن محققین را به این سمت سوق می‌دهد که بیش از اندازه به هماهنگی میان متغیرها اطمینان داشته باشند.

عدم پایایی[۱۵۲] مقیاس‌های اندازه‌گیری

اگر هر کدام از متغیرها به طور غیرپایایی محاسبه شده باشند، هرگونه نتیجه‌گیری در مورد کوواریانس میان متغیرها می‌تواند نادرست باشد (Nunnally & Bernstein, 1994). ناپایایی همواره می‌تواند همبستگی‌های دوگانه را تضعیف نماید. هنگامی که روابط سه یا بیشتر متغیر را شامل شود، اثر ناپایایی به میزان کمتری قابل پیش‌بینی است. نتایج مطالعه مکسول و دلانی (Maxwell & Delaney, 1990) نشان می‌دهد که عدم پایایی یک کوواریانس در یک تحلیل می‌تواند باعث تولید اثر مداخله معنادار شود، در حالی که اثر در واقع صفر است، و یا بالعکس اثر صفر تولید کنند، در حالی که در واقع اثر وجود داشته است. به همین طریق، رگوزا (Rogosa, 1980) پیشنهاد می‌کند اثرات پایایی در برخی طرح‌های همبستگی به الگوی روابط میان متغیرها و سطوح پایایی مختلف متغیرها وابسته است؛ به این صورت که، صرفنظر از اینکه اثر حقیقی چه بوده است، هر اثری یا اثر خنثی (صفر) می‌تواند یافت شود. در مطالعات طولی که نرخ تغییرات، تسهیل و یا دیگر مؤلفه‌های رشد را اندازه‌گیری می‌کنند، مشکلات ویژه‌ای می‌تواند در رابطه با پایایی بروز نماید (Willett, 1988). به همین دلیل پایایی باید برای هر یک از مقیاسها محاسبه و گزارش شود. از جمله راه‌حل‌های خلاصی از عدم پایایی می‌توان به افزایش تعداد اندازه‌گیری‌ها (مانند افزایش تعداد گویه‌های مقیاس و افزایش ارزیابها)، ارتقاء کیفیت اندازه‌گیریها (برای مثال،

149 Follow-up
150 ANOVA
151 MANOVA
152 Unreliability

گویه‌های بهتر و آموزش بهتر ارزیابها)، استفاده از نوع خاصی از تحلیل‌های منحنی رشد (Willett, 1988)، و استفاده از تکنیک‌هایی مانند مدلسازی متغیر مکنون بر روی مقیاس‌های مشاهده‌شده برای جدا کردن نمرات واقعی از واریاسیونهای خطا، اشاره کرد (Bentler, 1995).

محدودیت دامنه

برخی مواقع متغیرها به دامنه باریکی محدود می‌شوند. برای مثال، ممکن است در آزمایش‌ها دو مداخله بسیار شبیه به هم مورد مقایسه قرار بگیرند، و یا متغیر خروجی تنها دو مقدار داشته باشد، و یا اینکه مبتلا به اثرات سقف و کف باشیم. این محدودیت‌ها توان آماری را کاهش می‌دهند، و روابط دومتغیره را تضعیف می‌کنند. محدودیت موجود در متغیر مستقل را می‌توان از طریق مطالعه دوزهای کاملاً متفاوت مداخله، و یا دوز کامل در مقابل بدون‌مداخله کاهش داد. این کار خصوصاً در مراحل اولیه برنامه تحقیق، یعنی زمانی که احتمال وجود اثرات بزرگ تحت شرایط مناسب برای بروز آنها بررسی می‌شود، مفید و ارزشمند است. متغیرهای مستقل زمانی بواسطه اثر کف محدود می‌شوند، که تمام پاسخ‌دهندگان در حوالی پایین‌ترین نمره تجمع کنند. مانند زمانی که در یک مطالعه اغلب شرکت‌کنندگان در یک مقیاس اندازه‌گیری علائم بالینی افسردگی نمره نرمال می‌گیرند. از سوی دیگر زمانی که تمامی پاسخ‌دهندگان در حوالی بالاترین نمره مقیاس تجمع پیدا می‌کنند، اثر سقف محدودکننده متغیرهای مستقل خواهد بود؛ مانند زمانی که یک مطالعه محدود به باهوشترین دانش‌آموزان است. هنگامی که مقیاس‌های پیوسته تبدیل به مقیاس‌های دوتایی می‌شوند (یا سه تایی)، دامنه باز محدودتر می‌شود؛ مانند زمانی که محقق وزن میانه یک نمونه را برای ایجاد گروه‌های با وزن بالا و پایین مورد استفاده قرار می‌دهد. به طور کلی، بهتر است حتی‌المقدور از این تقسیم‌کردنها اجتناب کرد.[۱۵۳] آزمون پایلوت مقیاس‌ها و رویه‌های انتخاب می‌تواند برای تشخیص محدودیت دامنه کارساز باشد. همچنین اگر یک نمونه مناسب برای کالیبره کردن در دسترس باشد، تحلیل‌های نظریه پاسخ‌گویه[۱۵۴] می‌تواند برای حل این مشکل مفید باشد (Hambleton, Swaminathan, & Roger, 1991; Lord, 1980).

پایایی پایین شیوه اجرای مداخله

اگر مداخله در موقعیت‌های مختلف یا برای اشخاص مختلف به صورت غیرهمسان بکار گرفته شود، می‌تواند نتایج کوواریانس متغیرها را تحت تأثیر قرار دهد (Boruch & Gomez, 1977; Cook, Habib, Setterstern, Shagle, & Degirmencioglu, 1999; Lipsey, 1990). این تهدید در آزمایشات میدانی (فیلد) که در آنها کنترل مداخله

---

۱۵۳ بر خلاف انتظار، مطالعه مکسول و دلانی (۱۹۹۰) نشان داد که تبدیل مقیاس پیوسته دو متغیر به مقیاس دوتایی برای خلق یک طرح ANOVA فاکتوریل برخی مواقع می تواند توان آماری را افزایش دهد (از طریق افزایش نرخ خطای نوع اول).

154 Item Response Theory

دشوارتر از حالت آزمایشگاهی است، از اهمیت بیشتری برخوردار است. نبود اجرای استاندارد عموماً باعث کاهش اندازه اثر شده، و بنابراین ایجاب می‌کند توجه بیشتری به مؤلفه‌های طرح آزمایش که افزایش‌دهنده توان آزمون هستند (مانند اندازه نمونه)، مبذول شود. اگرچه برخی محققین بر این باورند که اجرای متفاوت مداخله برای شرایط یا افراد مختلف ممکن است نشان‌دهنده تعدیل مداخله برای افزایش اثر آن باشد (Scott & Sechrest, 1989; Sechrest, West, Phillips, Redner, & Yeaton, 1979; Yeaton & Secgrest, 1981). بعلاوه، اگر استنباط موردنظر درباره مداخله‌ای است که انتظار می‌رود در میان گروه‌های مختلف اثرات متفاوتی داشته باشد، استاندارد نکردن اجرای مداخله می‌تواند خالی از اشکال باشد. یقیناً، استاندارد نبودن، مشخصه ذاتی برخی مداخله‌ها در دنیای واقعی است. بنابراین، در مطالعات مرتبط با برنامه جامع ارتقاء کودکان و پیش دبستانی (Goodson, Layzer, St. Pierre, Bernstein & Lopez, 2000)، والدین بی‌بضاعت با توجه به نیازهای متفاوتی که داشته‌اند، بسته‌های کمکی متفاوتی دریافت کرده‌اند. در نتیجه، ترکیبی از آموزش‌های حین خدمت، آموزش‌های رسمی متداول، آموزش والدین، مشاوره و خانه‌های اورژانسی برای اسکان ممکن است مورد نیاز واقع شود؛ که این خود مداخله‌های ناهمگونی را در میان خانواده‌های مورد مطالعه ایجاد خواهد کرد. اگرچه، در تمامی این موارد باید تمام تلاش محقق معطوف به محاسبه عناصر بسته مداخله، و یافتن نحوه ارتباط این عناصر با تغییرات در نتایج باشد. از آنجا که این مسأله بسیار بااهمیت است، در فصل ۱۰ و ۱۲ روش‌های بهبود محاسبه و تحلیل اجرای مداخله که کاهش‌دهنده این تهدیدات است، را مورد بررسی قرار خواهیم داد.

واریانس‌های تصادفی و ناخواسته در شرایط آزمایش

اگر مؤلفه‌های یک مختصات آزمایشی به طور مصنوعی خطا را افزایش دهند، می توانند باعث مخدوش شدن کوواریانسهای حاصل شوند. از جمله این مؤلفه‌ها می‌توان به صداهای مزاحم و مختل‌کننده تمرکز، نوسانات در دمای محیط بواسطه سیستمهای سرمایش و گرمایش خراب، تغییرات مالی و اجرایی مکرر که مخلّ توجه شرکت‌کنندگان است، اشاره داشت. یک راه‌حل می‌تواند کنترل کردن این متغیرها باشد، یا انتخاب رویه‌های آزمایشی که توجه شرکت‌کنندگان را به مداخله جلب می‌کند. اما در بسیاری از شرایط فیلد، استفاده کامل از این راه‌ها غیرممکن است؛ در این شرایط لازم است آن منابع واریانسهای ناخواسته و تصادفی که قابل حذف یا کاهش نیستند را محاسبه کرده، و آنها را در تحلیلهای آماری متعاقب وارد نماییم. پایش کیفی اولیه آزمایش می‌تواند به محقق در شناسایی اینگونه متغیرها کمک نماید.

ناهمگونی پاسخ‌دهندگان

هر چه میزان ناهمگونی پاسخ‌دهندگان از نظر یک متغیر نتیجه‌ای در هر یک از شرایط آزمایش بیشتر باشد (واریانس درون‌گروهی)، انحراف معیارها روی آن متغیر (و همینطور متغیرهای همبسته آن متغیر) بالاتر خواهد

بود. اگر دیگر عوامل یکسان باشند، این ناهمگونی‌ها همبستگی (یا کوواریانس) سیستماتیک میان نتایج و مداخله را مبهم و نامعلوم می‌کند. همچنین زمانی که محققین از تعیین آن دسته مشخصات پاسخ‌دهندگان که می‌تواند در تعامل با رابطه علّی باشند، ناتوان هستند، احتمال بروز خطا افزایش می‌یابد. مانند برخی انواع افسردگی که باعث می‌شود فرد پاسخ بهتری به مداخله‌های روان‌درمانی بدهد. اینگونه ناهمگونیها نوعی خطا محسوب شده، و باعث مبهم شدن کوواریانس سیستماتیک می‌شوند، مگر اینکه به دقت شناسایی و در مدل لحاظ شوند. یکی از راه‌حل‌های موجود برای این مشکل آن است که نمونه‌هایی انتخاب شوند که از نظر متغیرهای دارای همبستگی، با متغیر نتیجه‌ای همگون و یکدست باشند. اگرچه، اینگونه شیوه‌های نمونه‌گیری اگر به درستی پایش نشوند، روایی بیرونی را کاهش داده، و دامنه را محدود می‌کنند. برخی اوقات بهتر است مشخصاتی که دارای همبستگی هستند را محاسبه کرده، و آنها را به عنوان متغیر بلوک‌بندی و یا متغیر تصادفی کمکی مورد استفاده قرار داد. همچنین می‌توان طرح‌های درون‌گروهی را بکار گرفت؛ در این طرح‌ها، میزان مزایا به اندازه همبستگی میان نمرات پیش‌آزمون و پس‌آزمون بستگی خواهد داشت.

تخمین نادرست اندازه اثر

هنگامی که اندازه اثر به طور ضعیفی محاسبه می‌شود، تخمین‌های کوواریانس می‌توانند نادرست باشند. برای مثال، هنگامی که مقادیر پرت موجب می‌شوند یک توزیع از توزیع نرمال دور شود، اندازه اثر به میزان قابل توجهی کاهش می‌یابد (Wilcox, 1995). ویلکاکس روش‌هایی را برای تخمین اندازه اثر در چنین داده هایی پیشنهاد می‌کند؛ اگرچه این راه‌حل‌ها ممکن است با تکنیک‌های استاندارد آماری چندان متناسب نباشد. همینطور، تحلیل متغیرهای نتیجه‌ای دوتایی (غیرپیوسته) با استفاده از مقیاس‌های پیوسته (ضریب همبستگی یا آماره تفاوت میانگین استاندارد شده) معمولاًموجب کوچک شدن اندازه اثر می‌شود (Fleiss, 1981, p.60). نرخهای فرد[۱۵۵] معمولاً گزینه بهتری به حساب می‌آیند. برای مثال، اگر یک آزمون t معمولی برای یک متغیر نتیجه‌ای دوقطبی (غیرپیوسته) استفاده شود، این تست آماری تفاضل میانگین استاندارد شده را بکار می‌گیرد؛ و توان آماری کمتری خواهد داشت. از آنجا که محققین به طور روزافزونی به گزارش کردن اندازه اثر و بازه اطمینان مقید هستند، هر روز موارد بیشتری از این دست عوامل که موجب تخمین نادرست اندازه اثر می شوند، کشف می‌شوند.

مسأله پذیرش‌فرض صفر

اگرچه در این کتاب سعی بر آن بوده که محققین را از این فکر که عدم رد فرض صفر به معنای نیافتن هر گونه اثر است، بر حذر داریم، اما برخی مواقع شرایطی بوجود می‌آید که باید چنین نتیجه‌ای گرفت. از جمله این

155 Odds ratios

شرایط، زمانی است که در آن فرض صحیح، عدم وجود اثر است؛ برای مثال اینکه اثرگذاری یک مداخله جدید هیچ تفاوتی نسبت به روش‌های پذیرفته‌شده ندارد، یا اینکه یک اثر جانبی خطرناک اتفاق نیافتاده است (Makuch & Simon, 1978)، آزمایش‌های فراحسی[۱۵۶] اثری نداشته است (Rosenthal, 1986)، و یا اینکه نتیجه پرتاب تاس برای بار اول هیچ ارتباطی با نتایج پرتاب‌های بعدی نداشته (یعنی سکه درست و منصفانه است) (Frick, 1996). از دیگر شرایط مطلوب‌بودن پذیرش عدم رد فرض صفر به عنوان عدم وجود اثر، آن است که یک سری از آزمایشات نتایجی بسیار نزدیک به هم بدست دهند، و محقق را به این باور برسانند که آیا واقعاً باید بررسی مداخله را ادامه دهد؟ سوم، موردیست که در آن محقق نشان می‌دهد گروه‌های مختلف از نظر تهدیدات مختلف روایی تفاوتی ندارد؛ مانند زمانی که هم‌ارزی گروه‌ها در مرحله پیش‌آزمون از نظر سوگیری انتخاب ارزیابی می‌شود (Yeaton & Sechrest, 1986). هر کدام از این موقعیت‌ها نیازمند آزمون این مسأله است که آیا کوواریانس بدست‌آمده به طور پایایی از صفر متفاوت بوده است؟ اگرچه اثبات اینکه کوواریانس برابر با صفر است، بسیار دشوار است؛ زیرا بر اساس نظریه توان آماری، حتی زمانی که اثر بسیار کوچک و ناچیز است، اگر اندازه نمونه بسیار بزرگ باشد، مقیاس‌های پایایی بالاتر بکار گارفته شوند، مداخله‌های بهتری و یا آماره‌های صحیح‌تری مورد استفاده قرار گیرند، می‌توان این اثرات را از صفر تمیز داد. براساس این نظریه است که ما نمی‌توانیم فرض صفر را تأیید کنیم (Frick, 1995).

برای مقابله با موقعیت‌هایی از این دست، ابتدا باید توان آماری را افزایش دهیم، به طوریکه نتایج بسیار نزدیک که قابل تمیز از یکدیگر نباشند، بوجود نیاید. جدول ۲.۳ فهرستی از راه‌های متنوعی که از آن طریق می‌توان توان آماری را افزایش داد را ارائه می‌نماید. اگرچه هر کدام از این راه‌ها از نظر توجیه‌پذیر بودن برای آزمایش‌های مختلف تفاوت دارند، و بعضی از این راه‌ها ممکن است در تناقض با دیگر اهداف آزمایش بوده و بنابراین مطلوب نباشند. با این وجود بررسی مطالعات از نظر این قبیل معیارهای توان مشخص خواهد کرد که آیا طراحی آزمایش جدید با توان بالا مطلوب و کاربردی خواهد بود یا نه.

دومین کاری که می‌توان انجام داد این است که توجه ویژه‌ای به تشخیص اندازه اثرهایی معطوف کنیم که ارزش دنبال کردن را داشته باشند؛ برای مثال، بالاترین حد آسیب قابل‌قبول، یا کوچکترین اثری که تفاوت کاربردی ایجاد می‌نماید (Fowler, 1985; Prentice & Miller, 1992; Rouanet, 1996; Serlin & Lapsley, 1993). مطالعه آشلفنتر (Aschenfelter, 1978) درباره اثر آموزش مهارت‌های انسانی بر درآمد افراد در آینده، چنین تخمین می‌زد که یک افزایش دویست دلاری در درآمدها برای نشان دادن موفقیت‌آمیز بودن برنامه کافیست. بنابراین محقق می‌توانست از تحلیل توان آماری به منظور اطمینان حاصل کردن از کافی بودن اندازه نمونه برای یافتن اثر استفاده کند. اگرچه، تعیین دقیق چنین اندازه اثری خود یک عمل سیاسی محسوب می‌شود؛ زیرا در این

156 Extrasensory

حالت یک نقطه مرجع درست شده است، که خلاقیت را می‌توان در قیاس با آن ارزیابی کرد. بنابراین حتی اگر خلاقیت اثر جزئی یا بخشی داشته باشد، اما نتوانسته باشد به سطح موردنظر اندازه اثر برسد، معتبر قلمداد نمی‌شود. از این رو مدیران برنامه آموزشی آموخته‌اند که به جای گفتن اینکه «می‌خواهیم دستاوردها را به اندازه دو سال به ازای هر سال تدریس افزایش دهیم»، بگویند «مایلیم دستاوردها را افزایش دهیم». اگرچه گزارش کردن مقدار دقیق اندازه اثر پس از انجام مداخله، این امکان را برای خوانندگان فراهم می‌آورد که تشخیص دهند آیا اثر کوچکتر از آن است که به آن توجه شود، یا اینکه به اندازه کافی قوی است که لازم باشد تحلیل‌های قویتری روی آن انجام شود.

سوم، آماردان‌های زیستی برای فرضیات مرتبط با هم‌ارزی دو مداخله، تکنیک‌های آزمون هم‌ارزی را تدوین کرده‌اند که می‌تواند به جای آزمون معناداری فرض صفر سنتی به کار گرفته شود. با استفاده از این روش‌ها می‌توان به بررسی این موضوع پرداخت که، آیا اثر مشاهده‌شده در دامنه‌ای که به قضاوت محقق به عنوان هم‌ارز برای اهداف کاربردی قابل قبول است، قرار گرفته یا نه، حتی اگر تفاوت میان مداخلهها صفر نباشد (Erbland, Deupree, & Niewoehner, 1999; Rogers, Howard, & Vessey, 1993; Westlake, 1988).

گزینه چهارم، استفاده از تحلیل‌های شبه‌آزمایشی است، برای بررسی اینکه آیا امکان وجود اثرات بزرگتر تحت برخی شرایط مهم -مثلاً بخشی از پاسخ‌دهندگان که به مداخله پاسخ قوی‌تری داده‌اند، یا واریته‌های دوزهایی که به طور طبیعی رخ داده‌اند و بالاتر از میانگین در یک آزمایش هستند- وجود دارد؟ در تفسیر این نتایج باید احتیاط شود زیرا این خطر وجود دارد که بر شانس تکیه کرده باشیم، و یا اینکه افراد غالباً به شیوه‌ای متفاوت مداخله‌ها را انتخاب کرده باشند (در واقع به انتخاب خود به مداخلهها تخصیص یافته باشند[157]). در هر حال، اگر تحلیل‌های شبه‌آزمایشی پیچیده نتوانند حداقل کوواریانس قابل‌توجه میان مداخله و نتایج را نشان دهند، اطمینان تحلیلگر نسبت به اینکه اثر کوچکتر از آن است که ارزش دنبال کردن داشته باشد، افزایش می‌یابد.

روایی درونی

اصطلاح روایی درونی عبارتست از استنباط‌هایی در باب اینکه آیا کوواریانس میان A و B نشان دهنده رابطه (اثر) علّی A بر B، به شکلی که متغیرها دستکاری و محاسبه شده‌اند بوده است یا خیر. برای نشان دادن صحت این استنباط لازم است تا محقق نشان دهد که A قبل از B اتفاق افتاده، A وB کوواریانس دارند، و اینکه هیچ توضیح یا تبیین منطقی دیگری برای رابطه مورد نظر وجود ندارد. مسأله اول به راحتی قابل‌حل است، زیرا در آزمایشها دستکاری A را قبل از B انجام می‌دهند. اگرچه ترتیب علّی تحقیقات غیرآزمایشی یک مشکل واقعی است؛ علی‌الخصوص در کارهایی که از نظر زمانی در مقطعی[158] هستند (مانند پیمایش).

157 self-selected
158 Cross-sectional

با وجود آنکه اصطلاح روایی درونی به طور گسترده‌ای در علوم اجتماعی بکار گرفته شده است، اما برخی استفاده‌کنندگان پایبندی چندانی به مفهوم روایی درونی که اولین بار توسط کمپبل (Campbell, 1957) ارائه شده نداشته‌اند. روایی درونی ربط چندانی به قابلیت بازتولید تحقیق (پایایی) (Cronbach, 1982)، قابلیت تعمیم به جامعه (Kleinbaum, Kupper, & Mrgenstern, 1982)، روایی محاسباتی (Menrad, 1991) و یا اینکه محقق ببیند آیا آنچه در واقع قصد بررسی آن را داشته است محاسبه کرده یا خیر، ندارد. کمپبل (Campbell, 1986) پیشنهاد می‌کند برای کاهش چنین سوء برداشتهایی بهتر است عنوان روایی درونی را به روایی علّی محلی ملکولی تغییر دهیم. تغییر نامی که اگرچه برای تبیین مفهوم سودمند است، اما دشواری بکار بردن آن باعث می‌شود تا محققین همان نام قبلی (یعنی روایی درونی) را ترجیح دهند. کلمه «علّی» در «روایی علّی محلی ملکولی» بر این نکته تأکید دارد که روایی درونی تنها به استنباطهای علّی مرتبط است و نه دیگر انواع استنباط‌هایی که در علوم اجتماعی انجام می‌شود. کلمه «محلی[۱۵۹]» به این مساله اشاره دارد که نتایج علّی محدود به زمینه خاص آن مداخله، نتایج، زمان، شرایط آزمون و افراد مورد مطالعه خاص است. کلمه «ملکولی» به این اشاره دارد که آزمایشات مداخله‌هایی را آزمون می‌کنند که خود بسته (مجموعه) پیچیده‌ای مشتمل بر عناصر متعددی هستند؛ تمامی این عناصر به عنوان یک کل واحد در شرایط مداخله آزمون می‌شوند. برای مثال، روان‌درمانی متشکل از مداخلات کلامی متفاوتی است که در زمان‌های مختلف و برای اهداف متفاوتی بکار گرفته می‌شوند. علاوه بر این، حرکات و اشارات غیرکلامی معمول در رفتار انسان، و به طور ویژه، در رفتار روان‌درمان با مراجعه کننده جزئی از مداخله به حساب می‌آید. دیگر عوامل، از دارونماهای تجویز شده توسط افراد دارای مجوز رسمی روان‌درمانی گرفته تا فضای اتاق روان‌درمانی و نوع بیمه پوشش دهنده خدمات روان‌درمانی همگی بخشی از بسته مجموعه مداخله به حساب می‌آیند. مراجعه‌کننده‌ای که به روان‌درمان مراجعه می‌کند، با تمامی این عوامل و نه تنها بخشی از عناصر که مطلوب محقق است، مواجه می‌شود. بنابراین استنباط علّی از یک آزمایش، به اثر تخصیص یافتن آزمایش‌شونده به تمامی اجزاء این بسته ملکولی اطلاق می‌شود. بدیهی است که آزمایش‌ها می‌توانند و در حقیقت باید این بسته‌های ملکولی را به بخش‌های کوچکتر آن که باید به طور انفرادی و یا در مقایسه با یکدیگر مورد آزمون قرار بگیرند، تقسیم نمایند. اما حتماً آن قطعات کوچکتر نیز خود بسته‌هایی هستند متشکل از اجزاء متعدد. بنابراین روایی درونی (که همان روایی علّی ملکولی است) به این موضوع می‌پردازد که آیا بسته‌بندی چندمتغیره و پیچیده مداخله موجب بروز تفاوتی در برخی متغیرها (آنطور که محاسبه شده‌اند) در یک زمان مشخص، با شرایط مشخص، و انواعی از افراد که مورد نمونه‌گیری واقع شده‌اند، شده است یا خیر.

159 Local

تهدیدات روایی درونی

مکی[۱۶۰] (Mackie, 1974) در باب این مسأله فلسفی که تحلیل علّی چیست می‌گوید: «به طور معمول، ما بواسطه حذف دیگر علل ممکن، از روی یک اثر، یک علّت را استنباط می‌کنیم» (ص ۶۷). تهدیدات روایی درونی همان دیگر علل ممکن هستند که می‌توانند در غیاب مداخله نیز رخ داده، و منجر به همان نتایجی شوند که در نتیجه اجرای مداخله، انتظار مشاهده آنها را داریم. جدول ۲.۴ فهرستی از این تهدیدها را به تفکیک ارائه می‌نماید؛ اگرچه که این تهدیدات مستقل از یکدیگر نیستند. این موارد برای تمامی انواع استنباط‌های علّی ملکولی کاربرد دارد، خواه این استنباط‌ها از آزمایش بدست آمده باشند، خواه در مطالعات همبستگی، مشاهده‌ای و یا مطالعات موردی. در نهایت، باید گفت که روایی دارایی یا مشخصه یک روش خاص نیست، بلکه مشخصه یک ادعای علمی است (Shadish, 1995b) (در اینجا ادعا در مورد دانش علّی).

جدول ۲.۴: تهدیدات روایی درونی: چرا این استنباط که رابطه میان دو متغیر علّی است، می‌تواند نادرست باشد؟

۱. ترتیب زمانی مبهم: عدم شفافیت درباره اینکه کدام متغیر ابتدا روی داده، می‌تواند باعث ایجاد ابهام درباره اینکه کدام متغیر علّت و کدام معلول بوده است، شود.
۲. انتخاب: تفاوت‌های سیستماتیک میان شرایط [مداخله و کنترل] از نظر مشخصات شرکت‌کنندگان که می‌تواند اثر مشاهده‌شده را بوجود بیاورد.
۳. گذشت زمان: رویدادها یا اتفاقاتی که همزمان با مداخله رخ داده، و می‌توانند دلیل بروز اثر مشاهده شده باشد.
۴. بلوغ: تغیراتی که به صورت طبیعی با گذشت زمان رخ می‌دهند، می‌تواند با اثر ناشی از مداخله اشتباه گرفته شود.
۵. رگرسیون: افرادی که به خاطر نمرات حادشان (بسیار بالا یا بسیار پایین) انتخاب می‌شوند، غالباً نمرات کمتر حادی در دیگر متغیرها می‌گیرند. این اتفاق می‌تواند به عنوان اثر مداخله اشتباه گرفته شود.
۶. ریزش: کاهش و از دست دادن شرکت‌کنندگان در جریان اجرای مداخله، و یا در زمان اندازه‌گیری متغیرها می‌تواند اثراتی مصنوعی تولید کند، البته در صورتی که ریزش به طور سیستماتیک با مختصات آزمایش همبستگی داشته باشند.
۷. آزمون‌کردن: مواجهه با یک آزمون می‌تواند نمرات مواجهه‌های بعدی با آن آزمون را تحت تاثیر قرار داده، و این اثر می‌تواند با اثر مداخله اشتباه گرفته شود.
۸. ابزار: ماهیت مقیاس اندازه‌گیری متغیر می‌تواند در طول زمان، یا برای شرایط مختلف آزمون و کنترل تغییر کند، به صورتی که با اثر مداخله اشتباه گرفته شود.

160 Mackie

۹. اثرات اضافه‌شونده و تعاملی تهدیدهای روایی: اثر یک تهدید می‌تواند به اثر تهدیدی دیگر اضافه شود، و یا می‌تواند به سطح (مقدار) دیگر تهدیدات وابسته باشد.

ترتیب زمانی مبهم

علّت باید قبل از اثر اتفاق افتاده باشد. اما برخی مواقع (علی‌الخصوص در مطالعات همبستگی) مشخص نیست که A قبل از B اتفاق افتاده یا بالعکس. البته حتی در مطالعات همبستگی نیز برخی اوقات یک جهت از اثر غیرمنطقی به نظر می‌رسد (مثلاً افزایش در مصرف سوخت گرمایشی باعث کاهش دمای بیرون از منزل نمی‌شود). همچنین برخی مطالعات همبستگی طولی[۱۶۱] بوده و در بیش از یک زمان داده جمع‌آوری می‌شود. این کار اجازه می‌دهد تا بتوانیم تنها متغیرهایی را به عنوان علّت تحلیل کنیم که قبل از اثرات احتمالی به وقوع پیوسته‌اند. اگرچه، صرف اینکه A قبل از B رخ داده باشد باعث نمی‌شود بتوانیم ادعا کنیم A باعث B شده است؛ و دیگر شرط‌های علّیت نیز باید وجود داشته باشند.

برخی علّیت‌ها دوطرفه (متقابل) هستند، مانند اینکه رفتارهای مجرمانه باعث زندانی شدن می‌شود، که زندانی بودن خود منجر به رفتارهای مجرمانه می‌شود. یا عملکرد بالای تحصیلی موجب افزایش اعتماد به نفس می‌شود که این خود می‌تواند باعث سطح بالاتری از عملکرد تحصیلی شود. بخش عمده این کتاب به آزمون روابط علّی یک طرفه می‌پردازد. در واقع آزمایش‌ها دقیقاً به منظور بررسی این نوع روابط طراحی شده‌اند؛ اینکه کدام عامل دستکاری‌شده قبل از اینکه دیگری اندازه‌گیری شود. با این وجود، می‌توان در آزمایش‌های مجزا بررسی کرد که آیا A موجب B شده، و یا B موجب A شده است. در نتیجه، آزمایشها بی‌ارتباط با اثرمتقابل علّی نیستند. دیگر روش‌های بررسی علّیت متقابل به طور خلاصه در فصل ۱۲ آمده است.

انتخاب

برخی اوقات در ابتدای آزمایش میانگین افراد شرکت‌کننده در شرایط آزمایش با میانگین افراد شرکت‌کننده در شرایط کنترل متفاوت است. این تفاوت می‌تواند دلیل هر نوع نتیجه‌ای باشد که در انتهای آزمایش بدست می‌آید و تحلیل‌کننده تمایل دارد آن را به مداخله نسبت دهد. فرض کنید یک برنامه آموزشی جبرانی به دانش‌آموزانی داده شود که والدین آنها ایشان را داوطلب کرده‌اند؛ و گروه مقابل شامل دانش‌آموزانی باشدکه داوطلب شرکت نبوده‌اند. والدینی که کودکان خود را برای برنامه داوطلب می‌کنند احتمال دارد به میزان بیشتری برای کودکان خود کتاب بخوانند، تعداد بیشتری کتاب در خانه داشته باشند، و یا در قیاس با گروه کنترل تفاوت‌هایی داشته باشند که دستاوردهای تحصیلی کودکان را تحت تاثیر قرار دهد. بنابراین کودکان شرکت‌کننده در برنامه جبرانی

161 Longitudinal

آموزشی ممکن بود بدون شرکت در دوره نیز عملکرد تحصیلی بهتری داشته باشند[۱۶۲]. هنگامی که تخصیص تصادفی به درستی انجام شود، اینگونه سوگیری‌های ناشی از انتخاب از بین می‌روند؛ چون تفاوت‌های میان گروه‌های بدست آمده از تخصیص تصادفی تنها به دلیل عوامل شانسی خواهد بود. بدیهی است که انجام نادرست تخصیص تصادفی خود موجب بروز سوگیری انتخاب می‌شود. همینطور که در آزمایشی که در آن اجرای تخصیص تصادفی بدرستی انجام شده اما ریزش نامتوازن در گروههای مداخله و کنترل موجب بروز تفاوت میان گروه‌ها می‌شود. سوگیری انتخاب در شبه‌آزمایش‌ها به طور وسیعی اتفاق می‌افتد. زیرا این نوع از مطالعات مشخصات ساختاری آزمایش را دارند، اما تخصیص تصادفی ندارند. مشخصه اصلی سوگیری انتخاب مخلوط شدن اثرات مداخله با تفاوت‌های ذاتی جمعیت‌های نمونه‌گیری است. بخش اعظمی از این کتاب به سوگیری انتخاب، چه زمانی که افراد خودشان خود را به گروه‌ها نسبت می‌دهند (خودشان گروهشان را انتخاب می‌کنند)، و چه در زمانی که محقق آنها را به گروه‌ها تخصیص می‌دهد، می‌پردازد.

گذشت زمان

این اثر تمام رویدادهایی را که در فاصله زمانی میان شروع مداخله و انجام پس‌آزمون رخ داده، و باعث می‌شوند تا نتایج مشاهده شده (حتی در غیاب مداخله نیز) ایجاد شوند، را در بر می‌گیرد. مثالی از تهدید گذشت زمان پیش‌تر در بحث مرتبط با ارزیابی برنامه‌های ارتقاء سلامت بارداری مورد اشاره قرار گرفت. در این مطالعه دریافت غذای کمکی به مثابه تهدید عمل می کرد (Shadish & Reis, 1984). در تحقیقات آزمایشگاهی، تهدید گذشت زمان را به وسیله جدا کردن پاسخ‌دهندگان از وقایع بیرون از آزمایشگاه و یا از طریق انتخاب متغیرهای وابسته‌ای که به ندرت احتمال دارد از عوامل دنیای بیرون تأثیر پذیرفته باشند (مثلاً یاد گرفتن سیلابهای بی‌معنا)، کنترل می‌کنند. اگرچه جدا کردن آزمایشگاهی در مطالعاتی که در محیط واقعی انجام می‌شوند، به ندرت امکانپذیر است. به این معنا که نمی‌توانیم مادران باردار را از دریافت کمک‌های غذایی و یا دیگر وقایع بیرونی که نتایج بارداری را تحت تأثیر قرار می‌دهد، منع نماییم. اگرچه حتی در تحقیقاتی میدانی نیز می‌توان منطقی‌بودن و معناداری خطر گذشت زمان را کاهش داد، مثلاً از طریق انتخاب گروه‌هایی از مکان یکسان، و یا از طریق تضمین این که زمان‌بندی آزمون برای هر دو گروه یکسان است (به این معنا که یک گروه در زمانی بسیار متفاوت از گروه دیگر آزمون نمی‌شود، مانند آزمایش کردن تمامی شرکت‌کنندگان کنترل قبل از آزمون شرکت‌کنندگان گروه مداخله (Murray, 1998)).

بلوغ

---

۱۶۲ اگرچه مسأله انتخاب در طرحهای دو گروهی متداول است، اما اینگونه سوگیری های انتخاب می تواند در طرح های تک گروهی نیز زمانی که ترکیب گروه در طول زمان تغییر می کند، رخ دهد.

شـــرکت‌کنندگان در پروژه‌های تحقیقاتی تغییرات بســـیاری را به طور طبیعی (حتی در غیاب مداخله) تجربه می‌کنند، مانند پیر شـدن، گرسـنه شـدن، عاقل‌تر شـدن، قوی‌تر شـدن، و یا با تجربه‌تر شـدن. برای مثال، یکی از مسائل موجود در بررسی اثرات برنامه‌های آموزشی جبرانی آن است که رشد طبیعی شناختی در کودکان باعث می‌شـود که عملکرد شـناختی کودکان در طول زمان ارتقاء پیدا کند (هدفی که برنامه‌های جبرای آموزشـی به دنبال آن هســـتند). حتی در مطالعات با بازه زمانی کوتاه‌مدت نیز چنین فرایندهایی نوعی مشـــکل به حســـاب می‌آیند. برای مثال، در آزمایش یادگیری کلامی، شرکت‌کنندگان ممکن است به سرعت دچار خستگی شوند و عملکردشـان کاهش یابد. در سـطح جامعه یا گروه، بلوغ عبارتسـت از تغییراتی که در طول زمان در یک جامعه آماری رخ می‌دهد، و می‌تولند نتایج را تحت تاثیر قرار دهد (Rossi & Freeman, 1989). برای مثال، اگر اقتصـــاد محلی در حال رشد است، سطح اشتغال احتمالاً رشد خواهد کرد، حتی اگر برنامه افزایش اشتغال هیچ تأثیری واقعی نداشـته بوده باشـد. تهدیدات مرتبط با بلوغ را اغلب می‌توان از طریق تضـمین اینکه تمامی گروه‌ها از نظر سن تقریباً مشابه هستند، و در نتیجه مرحله بلوغشان تقریباً یکسان است، کاهش داد. همچنین از طریق تضمین اینکه شـرکت‌کنندگان از یک منطقه جغرافیایی بوده و روندهای اجتماعی به طور متفاوتی آنها را تحت تأثیر قرار نمی‌دهد، می‌توان تهدید بلوغ را کاهش داد (Murray, 1998).

مصنوعات رگرسیونی[۱۶۳]

برخی مواقع پاسـخ‌دهندگان به این دلیل برای یک مداخله انتخاب می‌شـوند که نمرات آنها در یک مقیاس بالا یا پایین بوده اسـت. این مسـأله اغلب در شـبه‌آزمایشـها که در آنها، مداخله‌ها بر روی افرادی اجرا می‌شـود که واجد برخی شـاخص‌ها یا نیازهای خاص هسـتند، روی می‌دهد. در افرادیکه بدلیل نمرات بالا در یک متغیر یا شـاخص انتخاب می‌شـوند، این گرایش وجود دارد که نمرات پایین‌تری در دیگر شـاخص‌ها بگیرند؛ که این شـامل نمرات آنها در آزمون مجدد در همان آزمون (شاخص اولیه) نیز می شـود (Campbell & Kenny, 1999). برای مثال، فردی که بالاترین نمرات را در آزمون اول در یک کلاس کســـب کرده اســـت، کمتر احتمال دارد که در آزمون دوم نیز بالاترین نمره را کسـب نماید. و افرادی که به خاطر اسـترس بالا در اولین جلسـه به روان‌درمانی مراجعه می‌کنند، به احتمال زیاد در جلسات بعدی استرس کمتری خواهند داشت، حتی اگر روان‌درمانی تأثیری نداشته باشد. این پدیده رگرسـیون (حرکت) به سـمت میانگین نام دارد، و به سـادگی به جای اثر مداخله اشـتباه گرفته می‌شـود (Campbell & Stanley, 1963; Furby, 1973; Lord, 1963; Galton, 1886). مثال نوعی این پدیده زمانی دیده می‌شود که افراد یک مداخله را به این خاطر دریافت می‌کنند که نمرات پیش‌آزمونی بســیار بالا یا بســیار پایین کســب کرده‌اند. در این موارد این گرایش وجود دارد که نمرات کسب شـده در پس‌آزمون کمتر حاد باشـند. برخی اوقات اثر رگرســـیون به صـــورت پس‌گرا از نظر زمانی[۱۶۴] نیز اتفاق می‌افتد. به این معنی که هنگامی که افراد به دلیل نمرات پس‌آزمون بیشـینه یا حاد خود انتخاب می‌شـوند، نمرات پیش‌آزمون آنها کمتر حاد خواهد بود، و این روی

163 Regression artifacts
164 Retrospective

همان مقیاس‌ها اتفاق می‌افتد؛ درست مانند زمانی که مشاهدات حاد در یک پس‌آزمون، منتهی به مشاهداتی کمتر حاد در پس‌آزمونی همبسته می‌شود. به عنوان یک قانون کلی، خوانندگان باید وجود این تهدید را در مواقعی که پاسخ دهندگان به خاطر نمرات بالاتر یا پایین‌تر از میانگین‌شان انتخاب می‌شوند، بررسی نمایند.

رگرسیون به سمت میانگین به این دلیل اتفاق می‌افتد که مقیاس‌ها با یکدیگر کاملاً همبستگی ندارند (Campbell & Kenny, 1999; Nesselroade, Stigler, & Baltes, 1980; Rogosa, 1988). بخشی از توضیح برای چرایی وجود این همبستگی غیرکامل، خطای محاسبه تصادفی است. نظریه آزمون فرض می‌کند که هر مقیاس یک جزء نمره‌ای صحیح دارد که نشان‌دهنده یک قابلیت صحیح است، مانند افسردگی یا ظرفیت برای کار؛ و یک جزء خطا که به طور نرمال و به صورت تصادفی در اطراف میانگین آن مقیاس توزیع شده است. در هر موقعیت مفروض، نمرات بالا دارای خطاهای تصادفی بیشتری هستند که آنها را به سمت بالا می‌کشد. در همان مقیاس در زمانی دیگر، و یا در مقیاس دیگری در همان زمان، خطای تصادفی کمتر احتمال دارد به همان حد از شدت باشد، در نتیجه نمرات مشاهده‌شده (همان نمرات صحیح به علاوه خطاهای تصادفی کمتر حاد) کمتر حاد خواهند بود. در نتیجه بکارگیری مقیاس‌های پایاتر برای کاهش تهدید رگرسیون مفید خواهد بود.

با این وجود بکارگیری این مقیاسها از سوگیری رگرسیون جلوگیری نمی‌کنند، زیرا بیشتر متغیرها به طور ناکاملی با یکدیگر همبستگی دارند (به طور ماهیتی) و همچنان غیرهمبسته باقی خواهند ماند؛ حتی اگر به طور عالی و کامل مورد محاسبه قرار گیرند (Campbell & Kenny, 1999). برای مثال، محاسبه قد و وزن به طور یقین و کامل قابل‌انجام است؛ اما در هر نمونه، بلندترین فرد همواره سنگین‌ترین نیست و سبکترین فرد همواره کوتاهترین نیست. این نیز رگرسیون به سمت میانگین محسوب می‌شود. حتی زمانی که یک متغیر مشترک به طور دقیق و کامل در دو زمان متفاوت اندازه‌گیری می‌شود، مجموعه‌ای واقعی از نیروها می‌توانند موجب ایجاد نمرات حاد در یکی از این زمان‌های محاسبه شوند؛ نیروهایی که کمتر احتمال دارد در طول زمان پایدار بمانند. به عنوان نمونه، وزن یک فرد بالغ غالباً با خطای اندکی محاسبه می‌شود. اگرچه احتمال دارد افراد به این دلیل برای اولین بار به کلینیک کنترل وزن مراجعه کرده باشند که وزن آنها به دلیل مسافرت و مشکلات خانوادگی و استرس‌های ناشی از آن به طور ناگهانی بالا رفته است. همزمان با کاهش یا قطع این عوامل افزایش‌دهنده وزن این افراد گرایش به پایین آمدن خواهد داشت؛ حتی اگر مداخله کنترل وزن هیچ تأثیر واقعی نداشته باشد. اما باید به یاد داشت که در تمامی این موارد، نکته کلیدی برای امکان وجود مصنوعات رگرسیونی، که همان انتخاب بر اساس نمرات حاد است (خواه فردی باشد که بالاترین نمره را در اولین آزمون آورده و خواه فردی که در نمرات روان‌درمانی در سطح بالاتری قرار داشته و یا بلندترین فرد) وجود داشته است.

حال سوال این است که برای یافتن و کاهش سوگیری رگرسیون به سمت میانگین، محققین چه باید بکنند؟ اگر وجود نمرات حاد یکی از الزامات تحقیق است، بهترین راه‌حل آن است که گروه بزرگی از افراد با نمرات حاد را بوجود بیاوریم که از درون آن بتوان بوسیله تخصیص تصادفی افراد به گروه‌های مداخله انجام شود. این کار اثر رگرسیون به سمت میانگین را از اثر مداخله جدا کرده، و رگرسیون در تمام گروه‌ها به میزان یکسان اتفاق

می‌افتد. در مقابل، بدترین حالت زمانی رخ می‌دهد که شـــرکت‌کنندگان بر اســـاس نمرات حادشـــان در برخی متغیرهای غیرقابل‌اتکاء یا غیرپایا به گروهها تخصـــیص داده می‌شـــوند و ســـپس گروه حاصـــله با گروهی که به شـــیوه‌ای متفاوت انتخاب شـــده‌اند مقایســـه می‌شـــود. از آنجا که رگرســـیون زمانی واضـــح خواهد بود که نمرات اسـتاندارد شـده (در مقایسـه با نمرات خام) مورد بررسـی قرار می‌گیرد (Campbell & Erlebacher, 1970). در چنین مواردی باید آزمون‌های آسیب‌شناسی[۱۶۵] رگرسیون روی نمرات استانداردشده انجام شود (Galton; Campbell & Kenney, 1999). همچنین محقق باید پایایی و قابلیت‌اعتماد متغیرهای مبنای انتخاب اعضـــای گروه را به چند روش ارتقاء دهند. اول، از طریق افزایش تعداد گویه‌های سازه، دوم، از طریق میانگین گرفتن از آنها در چند نقطه از زمان، و یا ســـوم، با بکارگیری معادله‌ای چندمتغیره از چندین متغیر به جای اســـتفاده از یک متغیر در فرایند انتخاب. از جمله دیگر رویه‌ها،کار کردن با سـه نقطه زمانی یا بیشـتر اسـت. مثلاً بر اسـاس زمان اول انتخاب برای گروهها را انجام می‌دهیم؛ مداخله را بعد از محاســـبه انجام‌شـــده در زمان دوم اعمال می‌کنیم؛ و ســـپس به جای مقایســـه زمان اول و ســـوم، تغییرات میان زمان دوم و زمان ســـوم را مورد بررســـی و مقایســـه قرار می‌دهیم (Nesselroade et al., 1980).

رگرسـیون تنها در تحلیل کمّی اتفاق نمی‌افتد. روانشـناسـان این سـوگیری را به عنوان نوعی خطا که می‌تواند در فرایندهای شـناختی روزمره نیز رخ دهد شـناسـایی کرده‌اند (Fischhoff, 1975; Gilovich, 1991; G. Smith, 1997; Tversky & Kahnman, 1974). روان‌درمانها از دیرباز به این نکته اشـــاره می‌کنند که بیماران زمانی به روان‌درمان مراجعه می‌کنند که شـدت بیماری روحی آنها بسـیار بالاسـت، و این سـطح از بیماری با گذشـت زمان بهبود پیدا می‌کند، حتی بدون انجام روان‌درمانی. روان‌درمان‌ها پدیده رگرســـیون به ســـمت میانگین را بهبود خودبخود می‌نامند. فرایند پیشـرفت محاسـبه شـده بیمار تا حدودی بازگشـت به عقب وی است، یعنی حرکت وی به سـمت میانگین پایدار فردی خودش، پس از آنکه شـوک ناگهانی (مانند مرگ، بیکاری، یا تغییر در زندگی خانوادگی) که آنها را به سـمت اسـتفاده از خدمات روان‌درمانی سـوق داده بود، فروکش کرد. بسـیاری از مشـاوران کسـب و کار از محل سـرمایه گذاری بر اثر رگرسـیون به سـمت میانگین گذران زندگی می‌کنند. این مشـاوران از شـرکت‌هایی که به طور مسـتمر عملکرد بد داشـته‌اند اجتناب می‌کنند، اما بر کسـب و کارهایی تمرکز می‌کنند که اخیراً دچار یک رکود با دلایل نامشخص شده‌اند.

ریزش

ریزش (که برخی اوقات از آن به عنوان مرگ و میر آزمایشـــی نیز یاد می شـــود)، به این واقعیت اشـــاره دارد که شـرکت‌کنندگان آزمایش برخی مواقع نمی‌توانند آزمایش را تا پایان ادامه دهند. اگر انواع مختلفی از افراد در یک موقعیت (در مقایســـه با موقعیت دیگر) مورد محاســـبه قرار گیرند، این تفاوت‌های فردی می‌توانند نتایجی در پس‌آزمون بوجود بیاورند که الزاماً به دلیل وجود مداخله نبوده اســـت. بنابراین، در یک آزمایش تصـــادفی که به مقایسه اثر خانواده‌درمانی با استفاده از گروه‌های مباحثه برای درمان معتادین به مواد مخدر می‌پرداخت، مشاهده

165 Diagnostic

شد که معتادین با علائم شدیدتر به میزان بیشتری از گروههای مباحثه ریزش می‌کنند (در مقایسه با گروههای خانواده درمانی). حال اگر نتایج آزمایش نشان دهد که عملکرد خانواده درمانی در مقایسه با گروه‌های مباحثه ضعیف‌تر بوده است، این اثر مربوط ریزش است، و نه ماهیت مداخله زیرا آنها که علائم شدیدتری از اعتیاد را داشتند به میزان بیشتری در مداخله خانواده‌درمانی باقی می‌مانند (Stanton & Shadish, 1997). بنابراین، ریزش زیرمجموعه‌ای از سوگیری انتخاب قلمداد می‌شود که پس از اعمال مداخله رخ می‌دهد، اما بر خلاف سوگیری انتخاب، ریزش متفاوت را نمی‌توان از طریق تخصیص تصادفی به شرایط مختلف کاهش داد.

آزمون

بعضی مواقع یکبار انجام یک آزمون می‌تولند بر نتایج انجام همان آزمون در دفعات بعدی اثر بگذارد. تمرین، آشنایی، و یا دیگر اشکال عکس‌العمل مکانیسمهای مرتبطی هستند که می‌توانند اثراتی شبیه اثر مداخله تولید کنند. برای مثال، وزن کردن فرد ممکن است باعث شود تا فرد برای کاهش وزن خود تلاش کند، تلاشی که در حالت عادی انجام نمی‌داد. و یا گرفتن یک پیش‌آزمون در مورد دانش معنی لغات می‌تواند باعث شود که فرد به سراغ خواندن رمان رفته، و دانش لغات خود را افزایش دهد. از سوی دیگر، بسیاری از مقیاس‌ها قابل فعال‌سازی به این شکل نیستند. مثلاً اندازه‌گیری قد فرد قابلیت این را ندارد که فرد قد خود ارتقاء دهد (برای مثال نگاه کنید به Webb, Campbell, Schwartz, & Sechrest, 1966; Webb, Campbell, Schwartz, Sechrest, & Grove, 1981). تکنیک‌هایی از قبیل نظریه پاسخ گویه[۱۶۶] را می‌توان برای کاهش اثر سوگیری آزمون بکار گرفت. بر اساس این نظریه، می‌توان از آزمون‌های متفاوتی که تخمین‌های مشابهی از تواناییها بدست می‌دهند به جای یکدیگر استفاده کرد (Lord, 1980). برخی مواقع می‌توان از طرح چهار گروهی سولومون استفاده کرد. در این طرح برخی افراد پیش‌آزمونی را دریافت می‌کنند، در حالی که دیگران از آن مستثنی می‌شوند (Braver & Braver, 1988; Duke, Ullman, & Stein, 1995; Solomon, 1949). این کار برای آن انجام می‌شود که ببینند آیا پیش‌آزمون اثرات مداخله متفاوتی دارد یا نه. تحقیقات موجود نشان می‌دهد که لازم است تا آزمون به دقت مورد توجه و بررسی قرار گیرد (Willson & Putnam, 1982)؛ البته در مورد طرح‌هایی که در آن فاصله زمانی میان آزمون‌ها طولانی است، این اثر چندان مشکل‌زا نخواهد بود (Menard, 1991).

ابزار

در طول زمان تغییراتی هر چند ناچیز در ابزارهای محاسبه می‌تواند بوجود بیاید، و این تغییرات می‌توانند اثری شبیه به اثر مداخله تولید نمایند. برای مثال، فنر روی یک دستگاه پرس پس از مدتی ضعیف‌تر می‌شود و باعث می‌شود وزنه زدن آسان‌تر شود. در نتیجه به طور مصنوعی زمان عکس‌العمل افزایش پیدا می‌کند. و یا مشاهده‌گر انسانی (آزمون‌گر) با گذشت زمان، و تمرین با تجربه‌تر شده و بنابراین نمرات پس‌آزمون را با دقت بیشتری نسبت به نمرات پیش‌آزمون گزارش می‌کند. خطای ابزار خصوصاً در مطالعات مرتبط با رشد کودکان مسأله‌ساز می

166 Item response

شـود. در این مطالعات، واحد اندازه‌گیری یا مقیاس‌ها برای گروه‌های متفاوت سـنی معنای یکسـانی ندارد. خطای ابزار با خطای آزمون تفاوت دارد. از آن جهت که خطای ابزار به تغییر در ابزار اشـــاره دارد، در حالی که خطای آزمون به تغییرات بوجود آمده در شـرکت‌کنندگان (Shoknkoff & Phillips, 2000). خطای ابزار بیشـتر در مطالعات طولی اهمیت می‌یابد. در این مطالعات شیوه بکارگیری مقیاس‌ها با گذشت زمان تغییر کرده (نگاه کنید به شکل ۶.۷ در فصـل ۶)، و یا معنای یک متغیر در طول مراحل مختلف زندگی تغییر می‌کند (Menrad, 1991). روش‌های مختلف بررســـی امکان وجود خطای ابزار در کتاب کانینگهام (Cunningham, 1991) و هورن (Horn, 1991) مورد بحث قرار گرفته اسـت. محققان باید از تغییر ابزار آزمون در طول مطالعه پرهیز کنند؛ اما اگر تغییر ابزار اجتناب ناپذیر شـود، محقق باید هر دو ابزار قدیمی و جدید را حفظ کند، تا بتواند اثرات و نتایج آنها را با یکدیگر مقایسـه نماید (Murray, 1998).

اثرات تجمعی و تعاملی تهدیدات

تهدیدات مرتبط با روایی به طور مسـتقل و جداگانه اثر نمی‌گذارند. بسـیاری از آنها به طور همزمان عمل می‌کنند و زمانی که این اتفاق رخ می‌دهد، میزان خالص سوگیری، به جهت و قدرت هر سوگیری، بعلاوه اینکه سوگیری‌ها به صـورت تجمعی یا تعاملی با یکدیگر ترکیب شـده باشـند، بسـتگی دارد. البته در دنیای واقعی تحقیقات علوم اجتماعی محاسبه تعداد خالص این سوگیری بسیار دشوار است. فرض بر این است که هرچه تهدیدات با یکدیگر سـازگارتر و همگن‌تر باشـند، احتمال اسـتنباطهای علّی نادرسـت نیز بیشـتر خواهد بود. برای مثال، هنگامی که گروههای غیرهم‌ارز آزمایشی در ابتدای آزمایش شکل می‌گیرند، و در مراحل مختلف آزمایش بالغ‌تر می‌شوند، اثر ترکیب بلوغ-انتخاب مجال بروز می یابد. یکی از نشـانه‌های بروز این حالت می‌تواند این باشـد که دانشـجویان با عملکرد بهتر، بورسـهای ملّی را دریافت می‌کنند، و این دانشـجویان بیشـتر احتمال دارد که مهارت‌های آکادمیک خود را با سـرعت بیشـتری ارتقاء دهند. عملکرد بهتر در ابتدا، و رشـد آکادمیک سـریعتر هر دو می‌توانند باعث افزایش مصنوعی اثر مثبت مشاهده‌شده در بورس های ملّی شود. به همین طریق، اگر گروههای غیرهم ارزی که در شـرایط موقعیتی متفاوتی قرار دارند، و هر کدام سـرگذشـت محلی متفاوتی را تجربه می‌کنند در آزمایش بکار گرفته شوند، احتمال بروز سوگیری تجمعی گذشت زمان-انتخاب افزایش می‌یابد.

تخمین روایی درونی در آزمایشهای تصادفی و شبه‌آزمایشها

تخصـیص تصـادفی خطای انتخاب را حذف کرده، و باعث می‌شـود تنها امکان بروز خطاهای شـانسـی وجود داشـته باشـد. تخصـیص تصـادفی همچنین اثر دیگر تهدیدات روایی را نیز کاهش می‌دهد. از آنجا که گروه‌ها به صـورت تصـــادفی شـــکل گرفته‌اند، انتظار می‌رود هرگونه تفاوت ابتدایی در گروه‌ها از نظر نرخ بلوغ، از نظر تجربه کردن رویدادهای ناشـی از گذشـت زمان، و از نظر مصـنوعات رگرسـیونی شـانسـی (و ناشـی از عوامل غیرنظام‌مند) باشـد. مادامیکه محقق آزمونها و ابزارهای یکســـانی را برای گروه‌های آزمایش و کنترل بکار می‌گیرد، اثر پیش‌آزمون و

ابزار به طور یکسان در میان گروه‌ها اتفاق می‌افتد. بنابراین برای کاهش سوگیریهایی مانند پیش‌آزمون و ابزار، تخصیص تصادفی و رفتار یکسان با گروهها کارساز خواهد بود.

با فرض وجود تخصیص تصادفی، تنها در دو حالت همچنان ممکن است برای انجام استنباط‌های علّی با مشکل مواجه شویم. یکی هنگامی که مقدار ریزش در میان گروههای آزمایش متفاوت باشد؛ در این حالت تفاوت مشاهده‌شده میان گروهها به دلیل ریزش است و نه مداخله. تکنیک‌هایی برای مقابله با این مشکل طراحی شده است (Angrist et al., 1996) که در فصل ۱۰ همین کتاب به آنها خواهیم پرداخت. دومین حالت زمانی بروز می‌کند که لازم باشد آزمونهای متفاوتی در هر دو گروه انجام شود. این حالت زمانی پیش می‌آید که هزینه یا میزان خستگی ناشی از انجام مداخله به حدی بالاست که محقق تصمیم می‌گیرد تا پیش‌آزمون را تنها بر روی گروه آزمون (و نه گروه کنترل) که سطح همکاری بیشتری داشته و در صورت دریافت مداخله همکاری بیشتری خواهند کرد، انجام دهد. محقق باید مطالعه را بطور مداوم پایش نماید تا هرگونه ریزش متفاوت در گروهها را در مراحل اولیه تشخیص داده و اجازه ندهد مشکل زیاد بزرگ شود. همچنین می‌بایست فرایند آزمون را در گروههای متفاوت تا آنجا که ممکن است به طور مشابهی انجام دهند.

در شبه‌آزمایشها وضعیت استنباطهای علّی مبهم‌تر و دشوارتر است زیرا تفاوت میان گروهها در این نوع مطالعات به طور نظام‌مندتری اتفاق می‌افتد. بنابراین برای کاهش تهدیدهای روایی باید بر دیگر راهها تکیه کرد. مهمترین گزینه، تغییر و تعدیل عناصر طراحی مطالعه است. برای مثال، مصنوعات رگرسیونی را می‌توان از طریق انتخاب نکردن واحدهای مداخله بر اساس مقادیر حاد کاهش داد؛ البته به شرطی که این کار در تناقض با اهداف سوال تحقیق نباشد. یا اگر جداسازی آزمایشی توجیه‌پذیر باشد، امکان کاهش تهدید گذشت زمان وجود دارد. اگر چه استفاده از این عناصر طراحی همواره امکان‌پذیر نبوده و استفاده از آنها منجر به تغییر ماهیت سوال تحقیق می‌شود. و به همین دلیل است که تخصیص تصادفی تا به این اندازه مطلوب محققین است.

گزینه دیگر برای مقابله با تهدیدهای روایی، شناسایی تمامی انواع تهدیدها و خنثی کردن هر کدام یکی پس از دیگریست. شناسایی همه تهدیدها همواره وابسته به زمینه مطالعه است؛ برای مثال، آنچه در یک زمینه مطالعه به عنوان یک تهدید مطرح است (برای مثال، معرفی برنامه خیابان کنجد در جریان آزمایش روی آموزش مکمل در سال ۱۹۷۰)، ممکن است در زمینه‌ای دیگر اساساً تهدید به حساب نیاید (تماشای برنامه خیابان کنجد ابزاری ناموجّه برای کاهش حاملگی‌های ناخواسته است). پس از آنکه یک تهدید شناسایی شد، مقدار آن را می‌توان از طریق روش‌های کمّی و محاسباتی، و یا کیفی و با استفاده از مصاحبه‌ها و مشاهدات، برآورد نمود. در هر دو این موارد، اثر احتمالی تهدید موردنظر باید با اثر متغیرهای مستقل (نتایج تحقیق) مقایسه شود تا ببینیم آیا جهت سوگیری موردنظر همجهت با نتایج قابل مشاهده تحقیق است یا نه. اگر چنین است، تهدید قابل‌توجه و موجّه خواهد بود، مانند مثال معرفی برنامه خیابان کنجد به جای برنامه‌های کمک آموزشی به منظور کمک به ارتقاء مهارتهای خواندن. در غیر این صورت، تهدید موردنظر غیرموجّه است؛ مانند این کشف که سالمترین مادران بیشتر احتمال دارد از گروه مداخله حذف شوند، اما همچنان گروه مداخله بهتر از گروه کنترل عمل می‌کند.

هنگامی که تهدیدات را به صورت کمّی محاسبه می‌کنیم، ممکن است از تعدیلهای آماری بسیار مدرن و جدید استفاده شود که این خالی از اشکال نیست، زیرا این تعدیلها همواره درست نیستند، و همچنین زیرا نمی‌توانیم مطمئن باشیم که تمامی تهدیدهای مرتبط با روایی را شناسایی کرده‌ایم. بنابراین، ارزیابی موجّه‌بودن تهدیدهای روایی درونی به این شیوه - در قیاس با طراحی آزمایشی (تخصیص تصادفی) و بسیاری از عناصر طراحی که در طول این کتاب به آنها خواهیم پرداخت- بسیار انرژی‌بر و کمتر قابل اطمینان خواهد بود.

رابطه میان روایی درونی و روایی نتایج آماری

این دو نوع از روایی در ارتباطی نزدیک با یکدیگر قرار دارند. هر دو آنها با عملیات مطالعه (و نه به سازه هایی که این مطالعه‌ها در پی اندازه‌گیری آنهاست) و رابطه میان مداخله و نتایج سروکار دارند. روایی نتایج آماری خطاهای رخ داده در جریان ارزیابی کوواریانس آماری را بررسی می‌کند، درحالی که روایی درونی به خطاهای مرتبط با استدلالهای علّی مربوط است. حتی زمانی که تمامی تحلیلهای آماری در یک مطالعه بی‌عیب و نقص هستند نیز، خطاهای ناشی از استدلال علّی می‌تواند منجر به نتایج علّی نادرست شود. بنابراین کوواریانس آماری اثبات‌کننده وجود علّیت نیست. در مقابل، هنگامی که مطالعه‌ای بدرستی در قالب یک آزمایش تصادفی اجرا می‌شود، خطاهای آماری همچنان می‌توانند وجود داشته باشند، و منجر به قضاوت‌های نادرست درباره معناداری آماری و تخمین نادرست اندازه اثر شوند. بنابراین، در آزمایشهای کمّی، روایی درونی تا حد زیادی به روایی نتایج آماری وابسته است.

اگرچه، لازم نیست حتماً آزمایشها از نظر نتایج یا مداخله بکار گرفته‌شده در آنها، کمّی باشند (Lewin, 1953; Lieberson, 1985; Mishler, 1990)؛ و حتی برخی از صاحبنظران آنالیز آماری داده‌های کمّی را مخرب می‌دانند (Skinner, 1961). مثال‌های متعددی از آزمایش‌های کیفی در فیزیک و حتی علوم اجتماعی وجود دارد (مانند Drake 1981; Hacking, 1983; Naylor, 1989; Schaffer, 1989). مثلاً آزمایش معروف غار روبر (Sherif et al., 1961)، غالباً به طور کیفی انجام شد. در آن مطالعه، پسران حاضر در یک کمپ تابستانی به دو دسته یازده نفره تقسیم شدند. در ابتدا همبستگی درون‌گروهی در هر یک از گروهها به طور جداگانه تقویت شد. سپس تعارض میان‌گروهی بین آنها ایجاد شد. در نهایت، با استفاده از یک مداخله برای تسهیل همکاری و تماس کاری در حین کار برای نیل به اهداف مشترک، تعارضات کاهش داده شد. بسیاری از داده‌های این آزمایش کیفی بودند. این شامل اثر مشهور مشاهده شده روی کاهش تعارضات میان‌گروهی نیز می‌شود. در این موارد، روایی درونی، هیچ وابستگی مستقیمی به روایی نتایج آماری ندارد، اگرچه همچنان ارزیابی وجود کوواریانس میان مداخله و اثر مشاهده شده ضروری است (البته این ارزیابی کیفی است).

البته با چنین منطقی بود که کمپبل (Campbell, 1975) ادعای خود در خصوص غیرممکن بودن استفاده از مطالعه های موردی برای بررسی استنباط‌های علّی را پس گرفت (Campbell & Stanley, 1963)؛ چون استدلالهای مرتبط با استنباطهای علّی کیفی است، و چون تمامی الزامات منطقی برای استنباط کردن به همان میزان که برای کارهای کمّی کاربرد دارد، برای مطالعه کیفی نیز قابل‌استفاده است. اگرچه، استنباطهای علّی بدست‌آمده از

مطالعات موردی تنها تحت شـــرایط محدودی معتبر هســـتند (مانند زمانی که تفکیک علّت از دیگر متغیرهای کمکی تصـــادفی توجیه‌پذیر باشـــد). نگارندگان نیز با این نظر موافقند که علّیت الزاماً نیازمند نتایج مقیاس‌بندی شده یا مداخله‌های کمّی نیست.

فصل ۳

روایی سازه و روایی بیرونی

Relationship: ۱. شرایط یا حقیقت حالت مرتبط بودن، متصل بودن یا همراه بودن؛ ۲. اتصال نسبی یا سببی، دوستی

Trade-off: (اسم) تاخت‌زنی، بده‌بستان، مبادله یک چیز در مقابل چیزی دیگر، علی‌الخصوص چشم‌پوشی از یک منفعت یا مزیت برای بدست آوردن مزیت یا منفعتی دیگر که مطلوب‌تر است.

Priority: از ریشه انگلیسی میانه Priorite و فرانسه قدیم Pririts، ۱. اولویت و تقدم، ساختاریافته بر اساس ترتیب اهمیت یا اضطراری بودن؛ ۲. حق تقدم ثبت‌شده ؛ ۳. چیزی که تقدم دارد یا از نظر زمانی زودتر رخ می‌دهد؛ ۴. چیزی که شایسته توجه با اولویت است.

در فصل حاضر، مبحث روایی را با بحث در خصوص روایی سازه، روایی بیرونی و عوامل تهدیدکننده هر کدام پی می‌گیریم؛ و در انتها، فصل را با بحثی مبسوط در باب روابط، بده‌بستانها و اولویت انواع مختلف روایی به پایان خواهیم برد.

روایی سازه

گزارش اخیر آکادمی ملّی علوم در خصوص پژوهش درباره تحقیقات در مورد دوران اولیه کودکی به طور مجمل انواع مشکلات همراه با روایی سازه را به تصویر می‌کشد:

در اندازه‌گیری قد انسان (یا وزن یا ظرفیت ششها)، اختلاف‌نظر اندکی در مورد معنای سازه مورد اندازه‌گیری و یا در مورد واحد اندازه‌گیری (سانتی‌متر، گرم، سانتی‌متر) وجود دارد. اما اندازه‌گیری رشد در حوزه روان‌شناسی (مانند دایره‌لغات، استدلال کمّی، حافظه کلامی، هماهنگی چشم و دست و خودکنترلی) مشکل‌تر است؛ و

اختلاف‌نظر بیشتری درباره تعریف این سازه‌ها وجود دارد. بخشی از این مسأله به دلیل آن است که اغلب، واحدهای اندازه‌گیری طبیعی برای اندازه‌گیری این مفاهیم وجود ندارد (Shonkoff & Phillips, 2000, p. 82-83). اینجاست که دو مشکل روایی سازه -درک سازه‌ها و ارزیابی آنها- هویدا می‌شود. در این فصل نشان خواهیم داد که چطور این مشکلات در هنگام تبیین و اندازه‌گیری افراد، مختصات، مداخلهها و نتایج بکار رفته در یک آزمایش روی می‌دهند.

دانشمندان مطالعات تجربی را با مواردی خاص از افراد، مداخلهها، نتایج و مختصات آزمایشی انجام می‌دهند. اما این موارد خاص اغلب از آن جهت موردتوجه هستند که می‌توانند به عنوان مقیاسهایی[۱۶۷] از سازه‌های کلی در نظر گرفته شوند. روایی سازه عبارتست از انجام استنباط بر مبنای مشخصات و مقتضیات یک نمونه، در مورد سازه‌های مرتبه بالاتری که آن مشخصات و مقتضات نمایندگی می‌کنند (باز می‌نمایانند). برای مثال، در باب افراد مورد مطالعه، ممکن است یک اقتصاددان علاقمند به مطالعه سازه کارگران بیکار از قشر آسیب‌پذیر باشد، اما نمونه افرادی که وی در واقع مورد مطالعه قرار داده، ممکن است افرادی بوده باشند که در طول شش ماهه پیش از آزمایش درآمدی پایین‌تر از خط فقر داشته‌اند، و یا افرادی که در برنامه‌های غذایی کمکی و رفاهی دولتی شرکت کرده‌اند. اگرچه این اقتصاددان دوست دارد فاصله میان سازه و عملیات اندازه‌گیری مطالعه اندک باشد، اما برخی اوقات تفاوتها و اختلافهایی پیش می‌آید. مانند حالتی در این مطالعه، که برخی از کارگران ماهر که به تازگی شغل خود را از دست داده‌اند (و ظاهراً معیار بیکاری موردنظر محقق را دارا هستند) در مطالعه وارد شوند، در حالی که در واقع [به معنای موردنظر محقق] قشر آسیب‌پذیر به حساب نمی‌آیند (Heckman, Ichimura, & Todd, 1997). مثالهای مشابهی از این دست در مورد مداخلهها، نتایج و مختصات آزمایش وجود دارد. مثلاً روان‌درمانها ندرتاً علاقه‌ای به دانستن پاسخ‌های داده شده به ۲۱ سوال مقیاس افسردگی بُک دارند، بلکه آنها مایلند بدانند که مراجع آنها افسرده است یا نه. و یا هنگامی که اقتصاددانان کشاورزی روشهای مزرعه‌داری را در کوههای اطلس مراکش بررسی می‌کنند، فی‌الواقع علاقمند به مطالعه کشاورزی دیم در کشورهای فقیر هستند. یا زمانی که پزشکان نرخهای مرگ و میر در ۵ سال را در بیماران مبتلا به سرطان بررسی می‌کنند، در واقع در حال بررسی مفهوم کلی‌تر ماندگاری[۱۶۸] بیمار هستند.

همانطور که این مثالها نشان می‌دهند، بدون سازه‌ها نمی‌توان پژوهش انجام داد. همانطور که اینشتن زمانی گفته بود: «فکر کردن بدون طبقه‌بندیها و مفاهیم، به همان اندازه غیرممکن است که نفس کشیدن در خلاء (Einstein, 1949, p. 637-674)». روایی سازه به سه دلیل دیگر نیز اهمیت دارد. اول، سازه‌ها ابزار اصلی برای ربط دادن عملیات آزمایش به نظریه‌های مبنای تحقیق، و همینطور به ادبیات و زبانیست که جامعه کاربران دانش حاصل از آزمایش، بکار می‌گیرند تا نتایج آن را برای مصارف کاربردی مربوطه مورد استفاده قرار دهند. هرچه

167 Measures
168 Survival

آزمایشها حاوی سازه‌های پرخطاتری باشند، ریسک بدست آمدن نتایج گمراه‌کننده از آنها (خواه از نظر تئوریک و خواه کاربردی) افزایش می‌یابد. دوم، نام سازه‌ها اغلب بار معنایی اقتصادی، سیاسی و اجتماعی دارند (Hopson, 2000). این سازه‌ها ادراکها را شکل می‌دهند، به بحثها چهارچوب می‌بخشند، و همراهی‌ها و نقدها را به خود جلب می‌کنند. برای مثال، اختلاف‌نظرهای نژادی میان ذینفعان مختلف یک دعوی حقوقی مرتبط با «محیط کاری خصمانه از نظر نژادی» را در نظر بگیرید: مثلاً در مورد اینکه این سازه به چه معناست، چطور باید اندازه‌گیری شود، و آیا این تعاریف در هر زمینه‌ای قابل استفاده است؟ سوم، خلق سازه‌های اصلی و دفاع از آنها یکی از وظایف اصلی و بنیادین محققین در هر علمی بشمار می‌آید؛ مثلاً در علوم فیزیکی این مفاهیم شامل ساختن جدول تناوبی عناصر، تشخیص ترکیب آب، و کشف ساختار ژنها می‌شود (Mark, 2000, p.150). اگرچه ساختن سازه‌ها در علوم اجتماعی به دلایلی که در ادامه مورد بحث قرار خواهیم داد بسیار دشوارتر است.

چرا استنباط در مورد سازه‌ها یک مسأله به شمار می‌آید؟

نامگذاری برای چیزها و موضوعات مشکلی کلیدی در تمامی علوم است؛ زیرا نامها نشان‌دهنده عضویت در طبقه‌ای از موضوعات است، که این خود نشان‌دهنده رابطه مفهوم موردنظر با دیگر مفاهیم، نظریه‌ها و کاربردهاست. این موضوع حتی برای نامگذاریهای ساده نیز حائذ اهمیت است. برای مثال، چندی پیش روزنامه‌ها گزارشی از بحثهای میان ستاره‌شناسان برای نامگذاری ۱۸ جرم آسمانی تازه کشف‌شده منتشر کردند. ستاره‌شناسان اسپانیایی که این اجرام را کشف کرده بودند آنها را سیاره نامگذاری کرده بودند، اما دیگر ستاره‌شناسان استفاده از این نام را برای این اجرام نامناسب می‌دانستند، زیرا معتقد بودند برخی مشخصات این اجرام با مشخصات مفروض برای سیاره‌ها همخوانی ندارد. از نظر منتقدان این اجرام را می‌بایست کوتوله‌های قهوه‌ای نامید. اسپانیاییها اینطور استدلال می‌کردند که این اجرام کوچکتر از آن هستند که کوتوله نامیده شوند و سردتر از آن هستند که بتوان آنها را جوان در نظر گرفت. البته بدست آوردن روایی سازه در آزمایشهای میدانی[۱۶۹] کاری بسیار دشوارتر از آن چیزی است که در این مثال آورده شد.

برای داشتن روایی سازه چند کار باید انجام شود: (۱) باید در ابتدای کار تعریفی دقیق از افراد، مختصات آزمایشی، مداخلهها، و سازه‌های نتیجه‌ای موردنظر محقق ارائه شود؛ (۲) در هنگام انتخاب موارد (اعضای نمونه، اجزای مختصات و مداخله و غیره) دقت شود به گونه‌ای که مواردی انتخاب شوند که با تعاریف ذکر شده در شرط اول سازگار و منطبق باشند؛ (۳) میزان تطابق موارد انتخاب شده و سازه‌ها ارزیابی شود تا اطمینان حاصل شود که خطایی صورت نگرفته؛ و متعاقباً (۴) ویرایش و تعدیل توضیحات یا تعاریف سازه‌ها انجام شود. در این فصل ابتدا به موضوع تعریف و توضیح سازه پرداخته و برخی از رویه‌های معمول که موجب می‌شود محققین

169 Field

نمونه‌هایی را انتخاب کنند که بدرستی نماینده سازه‌های مورد مطالعه نیست، را مورد بحث قرار خواهیم. در بخشهای دیگر این کتاب، به بیان روشهایی که با استفاده از آنها می‌توان روایی سازه را بهبود بخشید خواهیم پرداخت. برای مثال فصل ۹ به این موضوع حساس اختصاص می‌یابد که چطور می‌توان اطمینان حاصل کرد که تعداد کافی شرکت‌کننده برای انجام آزمایش و تخصیص تصادفی وجود دارد؟ و فصل ۱۰ به این مسأله می‌پردازد که چطور می‌توان مداخله آزمایش را به درستی مفهوم‌سازی، اجرا و ارزیابی کرد.

مقادیر متنابهی از ادبیات موجود در حوزه فلسفه و علوم اجتماعی به مشکلات همراه با تبیین و توضیح سازه‌ها اختصاص یافته است (Lakoff, 1985; Medin, 1989; Rosch, 1978; Smith & Medin, 1981; Zadeh, 1987). احتمالاً نظریه‌ای که بیش از همه برای انجام این کار عمومیت‌یافته این است که هر سازه مشخصاتی دارد، مشخصاتی که برخی از آنها نسبت به برخی دیگر کلیدی‌تر محسوب می‌شوند. به عنوان مثالی ساده، درخت را فرض کنید. مشخصات مهم و نوعی (معمول) یک درخت عبارتند از اینکه بلند است، گیاهی چوبی با یک تنه مشخص و شاخه و برگهاست که حداقل ۳ سال عمر می‌کند. اگرچه هر یک از این مشخصه‌ها در هنگام بکارگیری با درجه‌ای از ابهام همراه خواهد بود. برای مثال، بلندی و تنه قطور قابل‌تشخیص، درخت را از بوته-که کوتاه‌تر است و شاخه‌های متعددی دارد- متمایز می‌کند. اما برخی درختها چند تنه دارند و برخی از درختها از برخی بوته‌ها کوتاه‌تر هستند. بنابراین هیچ مشخصه‌ای بنیادین نیست. در مقابل، این کتاب از منطق تطابق با الگو[۱۷۰] برای تصمیم در مورد اینکه آیا میزان کافی تطابق میان یک شیء یا فرد و مشخصه‌های نوعی یک طبقه یا دسته وجود دارد که بتوان عنوان آن طبقه را به شیء مورد نظر نسبت داد، استفاده می‌کند (خصوصاً با در نظر گرفتن عنوان طبقه‌های دیگری که می‌توانست به شیء موردنظر نسبت داده شود).

البته این مشخصه‌ها تنها شباهتهای سطحی هستند. دانشمندان غالباً به شباهتهای عمیقتری توجه می‌کنند. مشخصاتی اصلی که از نظر علمی اهمیت داشته، اما در نگاه یک فرد معمولی ممکن است سطحی و بی‌اهمیت جلوه کند. مثلاً برای یک فرد معمولی تفاوت میان درختهای برگریز و درختان همواره سبز، از نظر بصری اهمیت دارد. اما دانشمندان ترجیح می‌دهند درختان را به دو دسته گلدار (گیاهانی که گل می‌دهند، و در آنها دانه در یک محفظه محفوظ است) و بازدانگان (درختانی که گل نمی‌دهند و دانه‌های آنها در معرض محیط بیرون است، مانند کاج) تقسیم نمایند. دانشمندان این دسته‌بندی را نسبت به دسته‌بندی بر اساس برگریز بودن یا نبودن ترجیح می‌دهند؛ زیرا اطلاعات بیشتری در مورد فرایند تولید مثل درختان که در درک بهتر فرایند جنگلکاری و ماندگاری درختان اهمیت دارد، به دست می‌دهد. بنابراین تصمیم در مورد اینکه کدام مشخصات یک پدیده سطحی و کدام عمیقتر است دشوار است، اما محققین در هنگام انتخاب شرکت‌کنندگان، مختصات آزمایشی، مقیاسها و مداخلهها به طور ضمنی و یا آشکارا به این تصمیم‌ها دست می‌زنند.

170 Pattern-matching

دشواری از آنجا ناشی می‌شود که تصمیم‌گیری در این مورد که کدام مشخصه سازه موردنظر، مشخصه کلیدی است تا حدی وابسته به زمینه تحقیق است. برای مثال، نمی‌توان گفت که نظر دانشمندان صحیح است و نظر افراد عادی (در مورد انواع درختان) نادرست. برای فردی که قصد دارد خانه‌ای در مقابل یک محیط جنگلی پردرخت بخرد، برگریز بودن درختان به معنای هزینه فراوان سالانه برای جارو کردن برگهاست. مدین (Medin, 1989) مثال مشابهی می‌زند. او می‌پرسد چه نام یا برچسبی باید برای مجموعه‌ای که دربرگیرنده کودکان، پول، آلبومهای عکس، و حیوانات خانگی است انتخاب شود؟ وقتی معمولی نگاه می‌کنیم گزینه‌ای که بتوان گفت مشخصه اصلی مشترک این اعضا است را پیدا نمی‌کنیم. اما در یک زمینه یا زمان این کار شدنی است، و آن وقتی که قرار است تصمیم بگیریم در هنگام آتش‌سوزی کدامیک را باید نجات دهیم.

انتخاب ویژگی کلیدی، به ادبیات مورداستفاده جوامع مختلف علمی نیز بستگی دارد. کتاب «زن، آتش و خطر» نوشته لاکف (Lakoff, 1985) را در نظر بگیرید. اغلب ما به سختی می‌توانیم زن، آتش و خطر را در یک طبقه یا گروه دسته‌بندی کنیم. این سه در چه صفتی مسترک هستند؟ آیا زنان آتشی هستند؟ آیا زنان و آتش هر دو خطرناک هستند؟ در ادبیات جامعه دانشمندان علوم طبیعی آتش به طبقه‌ای مرتبط با اکسیداسیون قرار می‌گیرد، اما زن در آن دسته قرار نمی‌گیرد. در ادبیات جامع فسلفه باستانی، آتش یکی از سه عنصر پایه‌ای است که در کنار آب، باد، و زمین عناصر اربعه را تشکیل می‌دهند. ولی چیزهای خطرناک در میان این عناصر نیستند. اما در ادبیات کلامی علوم ماوراء طبیعی استرالیا که به آن دایربال[۱۷۱] گفته می‌شود، زن، آتش، و چیزهای خطرناک عناصری از یک دسته و طبقه محسوب می‌شوند.

تمامی دشواریهای ذکر شده برای انتخاب مهمترین مشخصه سازه‌ها، در علوم انسانی پررنگتر و غلیظتر نیز می‌شود؛ تا حدی به این دلیل که بسیاری از این سازه‌های مهم در حال کشف‌شدن و ساخته‌شدن هستند، بنابراین اجماع محکمی در مورد مشخصه‌های هر یک از این سازه‌ها وجود ندارد. در این شرایط عدم‌وجود اجماع، احتمال لغزش میان سازه و مصادیق عملیاتی آن در تحقیق افزایش می‌یابد. علّت دیگر، افزایش دشواری انتخاب عناصر سازه در علوم اجتماعی، ماهیت انتزاعی مسائلی است که علوم اجتماعی با آنها سروکار دارد (مسائلی مانند خشونت، انگیزه، تصمیم، برنامه و تمایل). این باعث می‌شود تا یکی از نظریه‌های طبقه‌بندی – نظریه نوع‌های طبیعی[۱۷۲] – که بطور گسترده در دیگر حوزه‌ها بکار گرفته می‌شود، در علوم اجتماعی بلااستفاده باشد. بر اساس این نظریه، طبیعت هر چیز را در محل زانوها و اتصلات بریده‌بریده کرده است، بنابراین ما نامها و درکهای مشترک خود را از روی موجودیتهایی که بواسطه زانوها و اتصالات از یکدیگر مجزا شده‌اند، شکل می‌دهیم. بنابراین، کلمات مجزایی برای تنه و شاخه‌های یک درخت داریم، اما کلمه‌ای برای قسمتهای تحتانی چپ یک

171 Dyirbal

172 The theory of natural kinds

درخت نداریم. در علوم اجتماعی تعداد بسیار کمتری زانو (یا اتصال معادل آن) وجود دارد؛ مثلاً جای زانوها در سازه‌هایی مانند خشونت یا تمایل کجاست؟

به دلیل وجود این مشکلات هیچگاه امکان برقراری یک رابطه یک-به-یک میان اجزاء عملیاتی یک مطالعه، و سازه‌های موردنظر وجود ندارد. پوزیتیوسیتهای منطقی[۱۷۳] به اشتباه فرض را بر این می‌گذارند که امکان انجام چنین کاری وجود دارد. از همین رو، حول عنوان عملیاتی‌سازی مبتنی بر تعریف[۱۷۴]، نوعی نظریه طراحی کرده‌اند؛ به این مضمون که، هر موجودیت عبارتست از آن چیزی که مقیاس اندازه‌گیری آن در بر می‌گیرد، بگونه‌ای که هر مقیاس انعکاس‌دهنده و نماینده‌ای تام و کامل از سازه خود است. عملیاتی‌سازی مبتنی بر تعریف به دلایل متعددی با شکست مواجه شده است (Bechtel, 1988; H.I.Brown, 1977). و یقیناً انواع مختلف عملیاتی‌سازی مبتنی بر تعریف، تهدیدی برای روایی سازه به شمار می‌آیند. بنابراین، نظریه مرتبط با سازه‌ها باید بر سه موضوع تاکید داشته باشد: (۱) هر سازه را در یک مطالعه و در میان مطالعات مختلف، به روشهای متعددی عملیاتی نماید؛ (۲) تطابق الگویی میان مشخصات چند-متغیره موارد نمونه (افراد، مشخصات، مداخله) با مشخصات سازه موردنظر را بررسی نماید؛ (۳) بحثهای درست و مشروع موجود در مورد کیفیت تطابق پیشگفت را بپذیرد، با توجه به اینکه ماهیت عملیات و سازه‌های موجود، هردو برساخته اجتماعی[۱۷۵] هستند. برای تسهیل انجام این سه مرحله باید تعریفی دقیق و همراه با جزئیات از موارد مطالعه شده داشته باشیم، توضیح و تبیین روشنی از عناصر کلیدی سازه موردنظر در اختیار باشد، و مشاهده‌ای واجد روایی از روابط میان موارد مشاهده‌شده سازه موردنظر، و هر سازه با اهمیت دیگر صورت گیرد.[۱۷۶]

ارزیابی خصوصیات نمونه‌گیری

تبیین و تشریح دقیق یک سازه برای دستیابی به روایی سازه ضروریست، اما این تنها نیمی از راه دستیابی به روواییست. نیمه دیگر، ارزیابی دقیق مشخصات نمونه‌گیری در یک مطالعه است، بطوریکه محقق بتواند تطبیق میان این ارزیابیها و سازه موردنظر را برآورد کند. برای مثال، بحث میان ستاره‌شناسان در مورد نامگذاری ۱۸ جرم یافته‌شده، مستلزم آن است که ابتدا مشخصات اصلی سیاره‌ها و کوتوله‌ها بیان شود، و سپس مشخصات این ۱۸ جرم تعیین‌شده، با مشخصات هر کدام از دسته‌ها (کوتوله‌ها و سیاره‌ها) مقایسه شود. از آنجا که مشخصات اصلی سیاره‌ها در میان ستاره‌شناسان کاملاً شناخته‌شده و پذیرفته‌شده است، منتقدین ابتدا به سراغ صحت و

173 Logical positivists

174 Definitional operationalism

175 Socially constructed

۱۷۶ کرونباخ و میل (۱۹۹۵) این مجموعه از روابط را شبکه نمولوژیک (nomological net) می نامند. البته در این کتاب از بکار بردن این اصطلاح پرهیز می‌کنیم زیرا این اصطلاح به لحاظ معنایی القاءکننده نوعی روابط قانونی است که به زعم نگارندگان، در مورد آزمایشها (آنطور که نگارندگان آنها را درک می‌کنند) مصداق ندارد.

درستی اندازه‌گیریها خواهند رفت. برای مثال، حدس می‌زنند که ستاره‌شناسان اسپانیایی درجه حرارت این اجرام را بدرستی اندازه‌گیری نکرده‌اند؛ در نتیجه، دیگر ستاره‌شناسان سعی می‌کنند اندازه‌گیریهای ایشان را تکرار کنند. برخی همان روش را مورد استفاده قرار می‌دهند و برخی دیگر روشهای دیگری را بکار می گیرند. اگر صحت و درستی اندازه‌گیریها تأیید شود، آنگاه باید مشخصات کلیدی سازه «سیاره» مورد بازنگری قرار گیرد، و یا شاید باید دسته‌ای جدید از اجرام آسمانی را تعریف کنیم تا در تطابق با اندازه‌های بدست آمده باشند.

بی‌تردید، این میزان توجه به اندازه‌گیری و مقیاس آن، نقشی بنیادین در شکل‌گیری مفهوم روایی سازه داشته (Cronbach & Meehl, 1955)، مفهومی که در نتیجه نیاز فزاینده آزمونهای روانشناسی برای دستیابی به کیفیت بالاتر بوجود آمد. از جمله وظایف تعیین شده انجمن روانشناسان آمریکا تعیین محاسبات و مقادیری است که باید پیش از انتشار یک آزمون مورد بررسی قرار گیرد. از میان مقادیر ضروری برای محاسبه، این انجمن به روایی سازه اشاره می‌کند. برای مثال، بر اساس نظر کرونباخ و میهل (Cronbach & Meehl, 1955) روایی سازه به این سؤال پاسخ می‌دهد که «چه سازه‌هایی مسئول واریانس مشاهده‌شده در عملکرد آزمون هستند؟» و همچنین اینکه «چطور باید بتوان از تفسیرهای پیشنهاد شده از آزمون دفاع کرد؟» (ص ۲۸۲). سازه‌ها و اندازه‌گیریهای حاصل از مقیاسها دو روی سکّه روایی سازه هستند.

به طور قطع، نظر کرونباخ و میهل (Cronbach & Meehel, 1995) اختصاصاً درباره آزمایشها ارائه نشده است. بلکه نظریه ایشان بیشتر در ارتباط با آزمونهای روانی‌ای که پس از جنگ جهانی دوم در روانشناسی کلینیکی رواج پیدا کرد، و به اندازه‌گیری سازه‌هایی مانند هوش، شخصیت، دستاوردهای علمی و یا آسیب‌شناسی روایی می‌پرداخت، ارائه شد. با این وجود، این آزمونهای روانی و سازه‌های مورد اشاره به طور وسیعی در آزمایشها نیز بکار گرفته شدند. بنابراین کاملاً طبیعی بود که نقدهای وارده به برخی یافته‌های تجربی خاص، روایی سازه استنباطهای صورت گرفته از طریق آن مقیاسها را زیر سوال ببرد. کوک و کمپبل (Cook & Campbell, 1979) روایی سازه را به سه گانه روایی ارائه شده توسط کمپبل و استنلی (Campbell & Stanley, 1963) افزودند؛ و همچنین کاربرد روایی سازه را گسترش دادند، به طوری که نه تنها نتایج، بلکه مداخله‌ها را نیز در بر می‌گیرد. آنها بر این باورند که تبیین دقیق مداخله و ماهیت آن، به اندازه تبیین سازه‌های اصلی تحقیق اهمیت دارد. در این کتاب، نگارندگان قدم را فراتر نهاده و تبیین افراد و مختصات شرایط آزمایش را نیز مشمول روایی سازه می‌دانند. یقیناً طبقه‌بندی آزمایشها به عنوان موجودیتهایی مشتمل بر افراد، مختصات آزمایش، مداخلهها، و نتایج یک دسته‌بندی، سلیقه‌ای و اختیاری از سوی نگارندگان بوده است، و آنها می‌توانستند مثلاً زمان را نیز به عنوان مشخصه‌ای مجزا در هر آزمایش در نظر بگیرند؛ همانطور که در کارهای پیشین این دو محقق این گونه رفتار شده است. بنابراین مفهوم روایی سازه در آزمایشها، عبارتست از انجام استنباط بر اساس ارزیابی هر یک از

مشخصات نمونه در یک مطالعه، در مورد سازه‌های سطح بالاتری[۱۷۷] که هر یک از آن مشخصات نمایندگی می‌کنند.

احتمالاً اغلب محققین منطق روایی سازه در مورد مطالعه مقیاسهای نتیجه‌ای[۱۷۸] تحقیق را درک کرده و می‌پذیرند. شاید بد نباشد با چند مثال، روایی سازه مرتبط با افراد، مختصات و مداخلهها را روشنتر کنیم. برخی از سازه‌های ساده‌تر در مورد تعریف مشخصات افراد واجد شرایط شرکت در آزمایش، نیازی به رویه‌های محاسباتی پیچیده ندارند؛ مانند زمانی که افراد را به زن و مرد تقسیم می‌کنیم. در این حالت هیچ بحثی در مورد اینکه آیا باید از خود-اظهاری [۱۷۹] استفاده کنیم یا مشاهده مستقیم، وجود ندارد. اما بسیاری از دیگر سازه‌ها که ممکن است برای تبیین افراد واجد شرایط بکار گرفته شوند، تا این اندازه مورد اجماع نبوده و محل مناقشه هستند. برای مثال، مسأله ساده و سطحی هویت نژادی و قومی افراد بومی شمال آمریکا را در نظر بگیرید. در طول سالیان، نامها و برچسبهای بکار گرفته‌شده برای اشاره به این گروه تغییر کرده است (هلندیها، آمریکاییهای مادری، افراد اول). محققین همچنین روشهای متنوعی را برای بررسی تعلق افراد به این گروه بررسی کرده‌اند؛ از گزارش شخصی گرفته (مثلاً با استفاده از فرمهای نظرسنجی در آمریکا)، تا ارزیابیهای رسمی درصد تباری مناسب (مثلاً با استفاده از ثبتهای قومی و قبیله‌ای مختلف). به همین ترتیب، میان افرادی که به عنوان بیمار اسکیزوفرنی تشخیص داده می‌شوند، تفاوتهای زیادی وجود دارد؛ بسته به اینکه تشخیص با استفاده از معیار انجمن رواندرمانی آمریکا (۱۹۹۴) انجام شده باشد، یا با یکی از ویرایش‌های اولیه همان دستورالعمل، یا بر مبنای آسیب‌شناسی ثبت‌شده در یک آسایشگاه، و یا طبق مقیاس اسکیزوفرنی مینه‌سوتا (Hathaway & McKinley, 1989). حتی زمانی که یک معیار عمومیت پیدا می‌کند نیز اگر برچسبهایی همچون آسیب‌پذیر را به طور سهوی و بی‌دقت بکار بگیرد (مانند آنچه در مثال مورد اشاره در ابتدای این فصل دیدیم)، جای تعجب نخواهد بود که افرادی با تفاوتهای فاحش همچنان تحت یک نام و برچسب خوانده شوند؛ نه تنها در مطالعات متفاوت بلکه حتی در دورن یک مطالعه.

سازه‌های بکار گرفته شده برای [تبیین و دسته‌بندی] *مختصات آزمایشی* نیز می‌توانند در طیفی از ساده تا پیچیده و متناقض قرار بگیرند. بسیاری از اوقات، مختصات آزمایشی مورد استفاده در تحقیق، نمونه‌ای در دسترس است که تحت عنوان «مرکز خدمات روانشناختی دپارتمانهای روانشناسی» معرفی می‌شود. محقق این مختصات را بر مبنای تجربه شخصی خود مورداستفاده قرار داده و غالباً هیچ اطلاعاتی در مورد اندازه مکان، شیوه‌های تأمین مالی، رفت آمد مراجعین و کارکنان و همچنین طیف بیماریهایی که تا به حال این مرکز با آن

177 Higher-order constructs
178 Outcome measures
179 Self-report

مواجه بوده ارائه نمی‌دهد. در حقیقت، این کلینیکها انواع مختلفی دارند، و برای مثال میان یک مرکز کوچک با تعداد محدودی مراجعه‌کننده که عموماً دانشجو هستند، و درمانگران آن غالباً از میان دانشجویان تحصیلات تکمیلی بوده، و تحت نظارت اساتید خود بیمار می‌بینند، با کلینیکهای بزرگی که تعداد زیادی درمانگر حرفه‌ای تمام وقت دارند، و این درمانگران طیف وسیعی از بیماریها را در مراجعان محلی درمان می‌کنند، تفاوت قابل‌توجهی وجود دارد. با این وجود، مختصات آزمایشی را می‌توان به طور رسمی‌تر و با کمک مقیاس طراحی‌شده توسط موس (مثلاً Moos, 1997)، و یا با کمک اطلاعاتی که از داده‌های تجربی قابل‌استنباط و استخراج است، ارزیابی کرد. مانند زمانی که تحلیل پروفایل مشخصه‌های خانه‌های پرستاری مبنای تعریف انواع مختلف خانه‌های پرستاری قرار می‌گیرد (برای مثال، Shadish, Straw, McSweeny, Koller, & Bootzin, 1981).

در خصوص روایی *مداخلهها*، بسیاری از حوزه‌های تحقیقاتی سنتها و رویه‌های کاملاً تعریف شده‌ای برای ارزیابی مشخصه‌های مداخلههایی که مورد استفاده قرار می‌دهند، دارند. مثلاً در مطالعات آزمایشگاهی روانشناسی اجتماعی فستینگر (Festinger, 1953) در مورد ناهماهنگی شناختی، سناریوهای مفصلی آماده شد تا اطمینان حاصل شود که مشخصه‌های اصلی ناهماهنگی شناختی در عملیات مطالعه لحاظ شده است. سپس این سناریوها با دقت تمرین و اجرا شد و در نهایت از محک مداخله[۱۸۰] استفاده شد تا اطمینان حاصل شود که شرکت‌کنندگان عملیات مطالعه را به صورت بازتابی (نماینده‌ای) دقیق از مفهوم موردنظر محقق درک کرده‌اند. این اندازه‌گیریها اطمینان محقق نسبت به این موضوع که مداخله در واقع به درستی اجرا شده است را افزایش می‌دهد. اگرچه اجرای محک مداخله برای برنامه‌های پیچیده اجتماعی مانند رواندرمانی یا سازماندهی مجدد کل مدرسه دشوار است. برای مثال در آزمایشهای رواندرمانی، آزمایش‌کنندگان اولیه عموماً تنها به گفتن نام درمانی که مورد استفاده قرار گرفته (مثلاً رفتاری، سیستماتیک، روان-دینامیک) اکتفا کرده اند. برخی مواقع این نامها با یک یا دو صفحه توضیح در مورد آنچه در درمان انجام شده و برخی اندازه‌گیری‌های صورت گرفته مانند تعداد جلسات همراه می‌شود. بکارگیری سیستم‌های پیچیده‌تر اندازه‌گیری درمان بیشتر حالت استثناء دارند تا قائده (Hill, O'Grady, & Elkin, 1992). این مسأله تا حدی به دلیل هزینه‌های انجام این اندازه‌گیریها و تا حدی به دلیل کمبود مقیاسهای مورد اجماع در خصوص بسیاری از درمانهاست.

نامگذاری نادرست *سازه‌ها* اغلب عواقب نامطلوب جدی، خواه به لحاظ نظری و خواه در عمل، در پی دارد. برای مثال، در مطالعه‌ای، افرادی با نمرات پایین در تستهای هوش را به عنوان «کند-ذهن» نامگذاری می‌کردند، اما بعدها مشخص شد که عملکرد پایین آنها می‌توانسته در اثر موانع زبانی و یا مواجهه ناکافی آنها با جنبه‌های فرهنگ آمریکایی که مبنای آزمون‌های هوش قرار گرفته است، بوده باشد. اثر ایجاد شده بر روی افراد، از

180 Manipulation-check

جابجایی در مدارس گرفته تا انگهای اجتماعی نسبت داده شده به آنها، بسیار وسیع بود. به همین ترتیب، برخی محققین رواندرمانی عنوان «درمانهای دارای پشتیبانی تجربی[۱۸۱]»(Chambless & Hollon, 1998; Kendall, 1998) را برای زیرمجموعه کوچکی از مداخلهها بکار می‌بردند. اینکار این تلقی را در میان محققین ایجاد می‌کرد که دیگر درمانهای روانشناختی دارای پشتوانه تجربی نیستند؛ علیرغم آنکه آزمایشهای انجام شده در طول چندین دهه اثربخشی این روشها را تأیید می‌کرد.

این نامگذاریهای غلط در توضیح یک آزمایش خواننده را در نحوه بکارگیری نتایج آزمایش در نظریه‌پردازیها یا عملیات موردنظر در تحقیق خود، گمراه می‌کند. یقیناً این یکی از دلایلی است که محققین کیفی تا این اندازه بر ارائه توضیحات مفصل در مورد اقدامات و اجزاء مختلف مطالعه تأکید دارند (به نحوی که خوانندگان بتوانند بجای تکیه بر نام‌گذاریهای انجام شده از سوی محقق، بر برداشتها و تعمیمهای طبیعی خود از نتایج و مشخصات مطالعه تکیه کنند). نگارندگان قویاً این رویکرد را تأیید می‌کنند، حداقل در محدوده دستورالعملهای گزارش‌دهی که عموماً در آزمایشها مورد رعایت قرار می‌گیرد. نگارندگان این کتاب همچنین اضافه کردن روشهای کیفی به آزمایشها را (برای افزایش این ظرفیت) مورد تأکید قرار می‌دهند.

مثالهای مورد اشاره نشان‌دهنده این مطلب هستند که ضرورتی ندارد ارزیابی مشخصه‌های مطالعات حتماً از طریق مقیاسهای چند-متغیره انجام شود (اگرچه اطلاعات بدست‌آمده از این مقیاسها اغلب مفیدتر است).

ارزیابی عبارتست از هر روشی که برای جمع‌آوری داده در مورد مشخصات نمونه‌گیری بکار گرفته می‌شود. مثلاً این روشها می‌تواند رکوردهای آرشیوی باشد؛ مانند جداول بیمارستانهای رواندرمانی. این جداول حاوی اطلاعات مرتبط با آسیب‌شناسیها و علائم بیماری بوده، و غالباً به صورت دستی نوشته می‌شوند، و پاسخ‌دهندگان هویتهای نژادی و قومی خود را با علامت زدن سوالات چند-جوابی نشان می‌دهند. این روشها می‌تواند همچنین شامل مشاهدات کیفی باشد. برخی اوقات مشاهده‌هایی رسمی مانند زیرنظر گرفتن شرکت‌کنندگان یا مصاحبه‌های باز توسط یک مردم‌شناس آموزش‌دیده انجام می‌شود، اما غالباً بصورت گزارش ساده محقق درباره مشخصات شرایط آزمایش بیان می‌شوند؛ مثل زمانی که محقق در جریان رفت و آمد به محل کار، مشاهدات خود درباره شرایط محیط را در قالب عبارت «همسایگی یک محل فقیر» بیان می‌کنند. ارزیابی‌ها حتی می‌تواند شامل مداخلههای آزمایشی‌ای باشد که برای روشنتر کردن ماهیت عملیات مطالعه طراحی شده‌اند؛ مانند زمانی که یک مداخله با یک دارونمای کنترل مقایسه می‌شود، برای اینکه مشخص شود یک مداخله تا چه اندازه مشابه دارونما عمل می‌کند.

یقیناً از گذشته تا به امروز، توجه به روایی سازه در آزمایشها به صورت نامتوازنی میان افراد، مداخله‌ها، نتایج و مختصات آزمایشی تقسیم شده است. توجه به اینکه *مختصات آزمایشی* به چه میزان توانسته انعکاس‌دهنده

181 Empirically supported psychological therapies

سازه‌های آزمایش باشد احتمالاً در اولیت پایینتری قرار دارد. مگر در تحقیقاتی که به نقد محیط و فرهنگ می‌پردازند. به همین ترتیب، در اغلب تحقیقات آزمایشی کاربردی نیز تمرکز بیشتر بر روایی سازه‌های نتیجه‌ای است. زیرا پرداختن به روایی مختصات آزمایش ممکن است بی‌معنا به نظر برسد، مگر در مواردی که محقق در حال بررسی مفاهیمی همچون بازگشت به رفتارهای خشونت‌بار، اشتغال، و یا موفقیتهای تحصیلیست. در تحقیقات پایه‌ای، توجه بیشتری به روایی سازه علّت معطوف می‌شود؛ به نحوی که ارتباط و اتصال آن با نظریه قوی باشد. این تفاوتها در اولویت‌بندی اجزاء روایی سازه تا حدی کارکردی است، و به نحوی شکل‌گرفته که نیازهای مختلف محققین در زمینه‌های مختلف تحقیقاتی را تأمین نماید؛ اگرچه بخشی از این فرایند شکل‌گیری نیز اتفاقی به نظر می‌رسد. اگر چنین باشد، افزایش توجه به روایی سازه افراد و مختصات آزمایش نیز سودمند خواهد بود.

در بخش بعدی افراد، مداخلهها، مختصات آزمایش و متغیرهای نتیجه‌ای را به طور مجزا مورد بحث قرار خواهیم داد. همانطور که در فصل اول نیز اشاره شد، عنوان و نام سازه‌ها می‌توانند برای بیان روابط میان عناصر یک مطالعه نیز بکار گرفته شوند. نامگذاری رابطه علّی میان مداخله و نتایج، یکی از مسائل مبتلا به در روایی سازه است. مانند زمانی که مداخله معینی را برای درمان سرطان به عنوان سیتو-توکسیک و یا سیتو-استاتیک نامگذاری می‌کنیم برای آنکه نشان دهیم درمان موردنظر تومور را مستقیماً می‌کشد، و یا اینکه با نامناسب کردن محیط پیرامون تومور، رشد آن را متوقف می‌کند. برخی دیگر نامها معانی مورد اجماعی را به خود گرفته‌اند، که در برگیرنده بیش از یک مشخصه یا معناست؛ مانند کلمه بیمه درمانی سالمندان[۱۸۲] در ایالات متحده که هم به معنای مداخله (خدمات درمانی) به کار می‌رود، و هم برای اشاره کردن به افراد دریافت‌کننده خدمات سالمندان (یا همان سالمندان).

تهدیدات روایی سازه

در ادامه فهرستی از تهدیدات احتمالی برای روایی سازه را به نمایش خواهیم گذارد. این تهدیدها تطابق میان عملیات مطالعه، و سازه‌هایی که برای تبیین و توضیح این عملیات بکار گرفته می‌شوند، را به خطر می‌اندازند. برخی اوقات مشکل به تبیین و توضیح سازه‌ها برمی‌گردد، و برخی دیگر از مواقع، مشکل از طراحی نمونه‌گیری و مقیاسهاست. عملیات یک مطالعه ممکن است نتواند تمامی مشخصات سازه‌های مرتبط را در برگیرد (به این مسأله نمایندگی ناقص از سازه گفته می‌شود)، و یا بالعکس، دربرگیرنده محتوایی مضاف بر سازه باشد. تهدایداتی که در ادامه به آنها خواهیم پرداخت، اشکال جزئی‌تری از این انواع کلی تهدیدات هستند، مواردی که به طور مکرر در تحقیقات و آزمایشهای گذشته رخ داده است. پنج تهدید اول مشخصاً برای افراد، مختصات آزمایشی،

182 Medicare

مداخلهها، و نتایج کاربرد دارد. مابقی تهدیدات بیشتر در ارتباط با روایی سازه نتایج و خصوصاً مداخلهها قرار دارند. این موارد غالباً از لیست ارائه شده توسط کوک و کمپبل (Cook & Campbell, 1979) اخذ شده است. نگارندگان می‌توانند تعداد قابل‌توجهی از دیگر تهدیدات محتمل را به این لیست اضافه کنند. برای مثال، در جدول ۴.۳ در فصل چهار تهدیداتی که اپیدمی‌شناسان برای مطالعات مورد-کنترل معرفی کرده‌اند مطرح خواهند شد. تهدیدات ارائه شده تحت عنوان «تعیین و انتخاب نمونه مطالعه» به طور ویژه به روایی سازه در مورد افراد و مختصات آزمایش می‌پردازند. در اینجا، برای جلوگیری از طولانی‌شدن لیست حاضر، از آوردن آن موارد اجتناب می‌کنیم.

تبیین ناکامل سازه

عدم‌تطابق میان عملیات و سازه‌ها می‌تواند بدلیل تحلیل ناکامل سازه تحت بررسی ایجاد شود. برای مثال، بسیاری از تعاریف پرخاشگری دربرگیرنده تمایل به آسیب‌رساندن به دیگری، و بروز نتایج آسیب‌زا است. به بیان دقیقتر، باید میان سه حالتی که در ادامه می‌آید تفاوت قائل شد: (۱) کبودی چشمی که توسط فردی روی صورت فرد دیگر ایجاد می‌شود به دلیل آن که این دو نفر در یک نقطه کور جاده با یکدیگر برخورد کرده‌اند؛ (۲)کبودی چشمی که فردی روی صورت دیگری بجا می‌گذارد چون می‌خواسته خوراکی او را بگیرد (خشونت ابزاری) و یا می‌خواسته به وی آسیب برساند؛ (۳) تهدید کلامی یک پسر بچه نسبت به بچه‌ای دیگر به این مضمون که پسر بچه دیگر باید خوراکی خود را به وی بدهد، در غیر اینصورت یک بادمجان پای چشمش می‌کارد. اگر تمایل و عمل فیزیکی خشونت هر دو بخشی از تعریف باشند، تنها مورد دوم از سه مورد ذکر شده در بالا مصداقی از خشونت به حساب می‌آید. تبیین دقیق وکامل سازه‌ها باعث می‌شود بتوانیم موارد تحت‌مطالعه را به دقت با تعریف ارائه شده در تبیینها تطبیق دهیم. همچنین این کار برای خوانندگان پژوهش نیز این امکان را فراهم می‌آورد که عملیات و فرایندهای انجام‌شده در مطالعات پیشین را به نقد بکشند. مواقعی که چندین تعریف صحیح و منطقی وجود دارد، منابع [یا ادبیات موجود] و همچنین میزان پذیرش آن تعریف در ادبیات جامعه علمی مربوطه، تعیین‌کننده جهت‌گیری تحقیق خواهد بود.

هرچقدر هم که سازه از ابتدا با دقت تبیین شده باشد، باز هم این تبیینها را می‌باید پس از انجام مطالعه[۱۸۳] مورد نقد قرار داد، زیرا بعضی اوقات خود نتایج بدست‌آمده از تحقیق نشان می‌دهد که سازه نیازمند بازتعریف است. برای مثال، بسیاری از محققین اثرات بازدارنده حکم زندان برای رانندگان مست را با احکام سبکتری مانند جرائم نقدی مقایسه کرده‌اند. پس از اینکه مطالعات متعدد نشان داد که حکم زندان موارد تکرار جرم را کاهش نداده است (Martin, Annan, & Forst, 1993)، محققین این سوال را مطرح کردند که آیا حکم زندان واقعاً «سنگینتر»

183 Post-study

از جریمه نقدی است؟ توجه داشته باشید که موضوع بحث در اینجا عدم یافتن اثر معنادار نیست، بلکه اینجا سؤال اصلی این است که آیا یافته‌ها را می‌توان به عنوان نتایج مقایسه مداخله سنگین‌تر با سبکتر به حساب آورد؟

مارک (Mark, 2000) نشان داد که محققین در هنگام تبیین سازه‌ها دچار چهار خطای معمول می‌شوند: (۱) سازه ممکن است بیش از حد کلی تعریف شود، مثلاً اینکه مداخله یک مطالعه را «رواندرمانی» نامگذاری کنیم؛ حتی اگر بیشتر مشخصه‌های آن به «تحقیق رواندرمانی» نزدیک باشد؛ (۲) ممکن است سازه را بیش از اندازه دقیق و جزئی تعریف کنیم، مانند هنگامی که بگوییم سطوح مشخصات ناشادی[۱۸۴] بیماران روانی در آسایشگاهها در واقع همان مشخصات بیماران روانی در هر کدام از دیگر محلهای نگهداری فقراست (Shadish, Silber, Orwin, & Bootzin, 1985)؛ (۳) ممکن است با سازه اشتباهی سر و کار داریم، مانند آنچه که در مورد نامگذاری مهاجران به ایالات متحده به عنوان کندذهن اتفاق افتاد، به دلیل آنکه این افراد نمره پایینی در آزمون هوش کسب کرده بودند، در حالی که آزمون صورت‌گرفته بیشتر نشان‌دهنده میزان آشنایی با فرهنگ و زبان آمریکا بود تا هوش؛ و (۴) عملیات یک مطالعه در واقع منعکس‌کننده دو یا چند سازه است، اما تنها با یک سازه تبیین‌شده. به عنوان نمونه، مقیاسهایی که بر اساس مشخصه‌ها یا صفتهای موردراندازه‌گیری نامگذاری می‌شوند، باید همچنین بر اساس روشهایی که برای محاسبه آن صفتها بکار می‌برند نیز نامگذاری شوند (مثلاً خود-اظهاری افسردگی). همانطور که این مثالها نشان داد، هر کدام از این خطاها می‌توانند در رابطه با هر یک از چهار مؤلفه مطالعه - افراد، مختصات آزمایش، مداخله ها و نتایج- رخ دهند.

اختلاط مفهومی [۱۸۵]

فرایند عملیات در یک آزمایش ندرتاً نماینده *خالص* سازه‌های موردنظر مطالعه است. مثالی را که در ابتدای این فصل در مورد افرادی که به آنها صفت «بیکار» اطلاق می‌شد در نظر بگیرید. ممکن است محقق این عنوان را برای افرادی بکار ببرد که دقیقاً نماینده مفاهیم تحقیق وی باشند (یعنی افرادی که درآمد خانوادگی آنها برای ۶ ماه قبل از زمان شروع مطالعه کمتر از خط فقر بوده، و یا کسانی که مشمول برنامه‌های حمایتی دولتی و یا تأمین غذای گرم برای فقرا بوده‌اند). با این وجود، امکان دارد این افراد به طور نامتوازن (یا دارای سوگیری) به یک طبقه قومی یا نژادی مانند آفریقایی-آمریکایی تبارها تعلق داشته باشند. این مشخصات ثانویه، بخشی از تعریف محقق از سازه بیکار نیست، اما به طور گریزناپذیری این ویژگیها با مشخصات موردنظر محقق ترکیب شده است.

184 Unhappiness
185 Construct confounding

سوگیری ناشی از تعریف عملیاتی منفرد

بسیاری از آزمایش‌ها تنها یک تعریف عملیاتی از یک سازه را در نظر می‌گیرند. از آنجا که اتکاء به یک تعریف عملیاتی برای سازه، از یک سو سازه را محدود کرده (به نوعی آن را دست کم می‌گیرید)، و از سوی دیگر می‌تواند باعث ورود عوامل و متغیرهای غیرمرتبط شود، روایی سازه در مطالعات متکی بر یک تعریف عملیاتی، نسبت به مطالعاتی که چندین تعریف عملیاتی از سازه ارائه می‌کنند، پایین‌تر است. در نظر گرفتن تعاریف عملیاتی متعدد غالباً هزینه چندانی ندارد، از این رو در علوم اجتماعی انجام این کار، به رویه‌ای ارجح تبدیل شده است. استفاده از افراد متفاوت و زمانهای متعدد نیز توصیه می‌شود. اما اغلب آزمایشها تنها یک یا دو مداخله به ازای هر مداخله بکار می‌گیرند، و تنها یک مختصات آزمایش را مورد استفاده قرار می‌دهند. علّت آن است که انجام مطالعه در چند شرایط آزمایش پرهزینه بوده، و افزایش تعداد مداخلهها، نیازمند اندازه‌های بسیار بزرگ نمونه خواهد بود (در غیر این صورت اندازه نمونه برای هر سلول آزمایش بسیار کوچک خواهد بود). با این وجود، جایگزینی برای روش تنوع بخشیدن به مداخلهها وجود ندارد. از این رو، اگر کسی تأثیر تخصص سخنگو یا رابط[۱۸۶] را مورد بررسی قرار می‌دهد، می‌تواند سه سخنگوی فرضی و ساختگی را در نظر بگیرد: یک استاد برجسته مرد از یک دانشگاه مشهور، یک زن دانشمند از یک مرکز تحقیقاتی معتبر، و یک خبرنگار علمی معروف از آلمان. واریانس بدست‌آمده در اثر تفاوت در سخنگوها مورد بررسی قرار می‌گیرد تا مشخص شود که آیا انواع مختلف منبع پیام اثرات متفاوتی بر پاسخ‌های بدست آمده دارند؟ اگر این اثر وجود داشته باشد، این فرضیه که تخصص منبع سازه‌ای منفرد[۱۸۷] است، را باید مورد بازنگری قرار داد. اما حتی اگر اندازه نمونه اجازه تجزیه و تحلیل نتایج برای هر یک از این سخنگوها را ندهد، داده‌های بدست‌آمده از این سه گروه را می‌توان ترکیب کرده، و سپس بررسی کرد که آیا تخصص، علیرغم تمامی منابع ناهمگونی موجود در هر یک از سه عملیات، اثربخش بوده است؟

سوگیری ناشی از روش منفرد

داشتن بیش از یک تعریف عملیاتی مفید است، اما اگر تمامی مداخلهها به شیوه‌ای مشابه به پاسخ‌دهندگان ارایه شوند، خود روش فی‌النفسه می‌تواند نتایج را تحت تأثیر قرار دهد. همین مطلب در مورد حالتی که تمام مقیاسهای اندازه‌گیری نتایج از یک ابزار مشابه جمع‌آوری پاسخها استفاده می‌کنند، و یا حالتی که تمام مشخصات افراد از جداول بیمارستانی استخراج می‌شود، صدق می‌کند. بنابراین در مثال فرضی قبلی، اگر نظر متخصصین به صورت مکتوب به پاسخ‌دهندگان ارائه شده باشد، درست‌تر آن است که نام مداخله را «متخصصین ارائه شده بصورت مکتوب» بگذاریم، تا روشن شود که برای ما معلوم نیست که اگر صدای متخصصین شنیده

186 Communicator
187 Single construct

می‌شد، و یا خود آنها دیده می‌شدند، همچنان همین نتایج بدست می‌آمد یا نه. به همین ترتیب، غالب مقیاسهای اندازه‌گیری نگرش بدون در نظر گرفتن اینکه آیا عبارتها یا گویه‌های مقیاس به صورت منفی یا مثبت جمله‌بندی شده‌اند، و یا اینکه می‌شد به جای کاغذ و قلم از ضبط صوت استفاده کرد، به پاسخ‌دهندگان عرضه می‌شوند. از این رو امکان دارد نتایج متفاوتی از ارائه سؤالات به صورت مکتوب، در مقایسه با حالت ضبط صدا بدست بیاید. و یا در حالت دوم، به واسطه جهت‌گیری مشابه همه سؤالات (همگی منفی یا همگی مثبت)، امکان بروز سوگیری پاسخ وجود داشته باشد.

اختلاط سازه‌ها با سطوح سازه‌ها

برخی اوقات محقق یک نتیجه‌گیری کلی در مورد یک سازه ارائه می‌کند، غافل از اینکه در حقیقت تنها بخشی از سطوح (اندازه‌های) هر یک از ابعاد آن سازه را مطالعه کرده است، و در صورتی که دیگر سطوح آن سازه مورد اندازه‌گیری قرار می‌گرفت، نتایج می‌توانست متفاوت باشد. به عنوان مثال، در مطالعاتی که در آنها گروه آزمون و کنترل با یکدیگر مقایسه می‌شوند، ممکن است مقدار مداخله به حدی پایین باشد که هیچ اثری مشاهده نشود، و این نتایج محقق را به این باور برساند که مداخله موردنظر اثری ندارد. اما نتیجه‌گیری درست این خواهد بود که مداخله موردنظر در سطوح پایین (مثلاً در دوز پایین دارو) اثری ندارد. یک راه برای حل این مشکل، آزمون کردن سطوح (مقادیر) مختلف مداخله است. شکل پیچیده‌تر این سوگیری زمانی اتفاق می‌افتد که اثر دو مداخله مقایسه می‌شوند، اما این دو مداخله به طور هم‌ارز و مشابهی عملیاتی نشده‌اند. در اینجا، محقق ممکن است به نادرست چنین نتیجه‌گیری کند که مداخله الف بهتر از مداخله ب عمل می‌کند؛ در حالی که نتیجه‌گیری درست آن بود که بگوید مداخله الف در سطح ۱ بهتر از مداخله ب در سطح صفر عمل می‌کند. مشابه این نوع سوگیری اختلاط سطوح می‌تواند برای انواع شرکت‌کنندگان، نتایج و شرایط آزمایش نیز رخ دهد. مثل هنگامی که مشخصات فردی محدودی (مانند محدوده سنی خاص)، یا مشخصات شرایط آزمایش خاصی (مثلا منحصراً مدارس دولتی) استفاده شده، اما هیچ اشاره‌ای به این موضوع در گزارش تحقیق نمی‌شود.

ساختار فاکتوریل حساس به مداخله[۱۸۸]

هنگام بحث در خصوص روایی در فصل گذشته اشاره شد که تغییرات ابزاری می‌توانند حتی در غیاب مداخله نیز رخ دهند. با این وجود، تغییرات در ابزار گاهی اوقات به دلیل مداخله رخ می‌دهند، مانند مواردی که در آن افراد بواسطه قرار گرفتن در معرض یک مداخله آموزشی (در مقایسه با افرادی که با مداخله مورد نظر مواجه نشده‌اند) یاد می‌گیرند که به آزمون از دیدگاه متفاوتی نگاه کنند. برای مثال، افرادی که مداخله دریافت نکرده‌اند ممکن

188 Treatment-sensitive factorial structure

است در یک آزمون سنجش نگرش، نسبت به نژادی دیگر به شکلی هماهنگ پاسخ دهند، که منجر به بدست آمدن یک آزمون تک-عاملی تعصب نژادی شود. اما کسانی که در معرض مداخله بوده‌اند، ممکن است پاسخهایی با ساختار عاملی پیچیده‌تر بدهند (برای مثال، من در آزارهای فیزیکی یا کلامی در مکالمات مشارکت نمی‌کنم، اما در حال حاضر مشاهده می‌کنم که جکهای قومیتی نوعی تبعیض محسوب می‌شود، چیزی که من تا پیش از این متوجه آن نشده بودم). این تغییر در ساختار عاملی خود فی‌النفسه بخشی از نتایج مداخله به شمار می‌آید. اما کمتر محققی به دنبال ساختارهای مختلف عاملی در میان گروهها، به عنوان یک نتیجه می‌گردد. هنگامی که نمرات تمامی گویه‌ها در قالب یک کل برای هر دو گروه جمع بسته می‌شود، این تجمیع می‌تواند باعث تبیین اشتباه از سازه در حال محاسبه شود، زیرا چنین فرض می‌کند که این سازه در میان گروهها مشابه و در نتیجه قابل‌قیاس است.

تغییرات واکنشی خود-اظهاری

آیکن و وست (Aiken & West, 1990) بحثی را در مورد نوعی مشکل اندازه‌گیری که در اثر کاربرد روش خود-ابرازی بوجود می‌آید مطرح می‌کنند. مشکلی که بواسطه آن، ساختار عاملی و سطح پاسخها می‌تواند تحت‌تأثیر این مسأله که آیا فرد به گروه آزمون تخصیص پیدا کرده یا گروه کنترل، قرار بگیرد (حتی پیش از آنکه مداخله اجرا شود). برای مثال، شرکت‌کنندگانی که خواهان مداخله هستند، ممکن است به گونه‌ای رفتار کنند که به نظر نیازمندتر و واجد شرایط‌تر به نظر برسند (بسته به اینکه کدامیک راحتتر ایشان را به مقصود می‌رساند). اما پس از تخصیص افراد به گروهها، این احتمال وجود دارد که انگیزه این افراد مشتاق کاهش پایان یابد، در حالی که دیگر افراد دچار این افت انگیزه نشوند. در نتیجه، تفاوتهای احتمالی در نمرات پس‌آزمون نه تنها نشان‌دهنده تغییرات در تظاهرات بالینی بیماری است، بلکه ناشی از تفاوتهای رخ داده در سطح انگیزه است، در حالی که محقق احتمالاً به اشتباه تمامی تفاوتهای مشاهده‌شده را به حساب تغییر در تظاهرات بالینی خواهد گذاشت. در همین راستا، براخت و گلس (Bracht & Glass, 1968) بر این باورند که حساسیت‌زایی در پس‌آزمون (بر خلاف پیش‌آزمون)، زمانی مجال بروز می‌یابد که پس‌آزمون شرکت‌کنندگان را نسبت به مداخلات قبلی که شرکت‌کنندگان دریافت کرده‌اند، حساس نموده، و در نتیجه، موجب ایجاد پاسخی از سوی شرکت‌کنندگان شود که در وضعیتی غیر از این، امکان بروز نداشت. از جمله راه‌حلها برای حل این مشکل می‌توان به (۱) استفاده از مقیاسهای بیرونی[۱۸۹] (و نه خود-ابرازی) که کمتر واکنشی هستند (Webb et al., 1966, 1981)؛ (۲) بکارگیری تکنیکهایی که مشوق پاسخ صحیح هستند (مانند مرز بوگوس[۱۹۰] که در آن مشارکت‌کنندگان توسط دستگاهی تحت‌نظر گرفته می‌شوند، و به دروغ به آنها گفته می‌شود که دستگاه می‌تواند پاسخهای نادرست آنها را تشخیص

189 External measures
190 Bogus pipeline

دهد) (Jones & Sigall, 1971; Roese & Jamieson, 1993)؛ (۳) عدم‌دسترسی افراد تخصیص‌دهنده به نمرات پیش‌آزمون شرکت‌کنندگان؛ و (۴) در نهایت، بکارگیری گروههای مرجع یا معیارهای رفتاری برای معیارگذاری[۱۹۱] به پاسخها، اشاره داشت.

واکنشی بودن[۱۹۲] نسبت به مشخصات آزمایش

انسان به طور فعال موقعیتهای اطراف خود را تفسیر می‌کند. این شامل شرایط پیرامونی مداخله آزمایش نیز می‌شود. در نتیجه، معنای مفهومی بسته ملکولی مداخله دربرگیرنده این واکنشها نیز هست. این واکنشها می‌تواند اشکال مختلفی داشته باشند. بر اساس نظر روزنوایگ (Rosenzweig, 1933) ممکن است شرکت‌کنندگان تلاش کنند موضوع مورد مطالعه محقق را حدس بزنند و سعی کنند پاسخ‌هایی ارائه کنند که برای محقق مطلوب باشد. اورنه (Orne, 1959, 1962, 1969) نشان می‌دهد که «مشخصات تقاضا» در شرایط موقعیتی آزمایش می‌تواند برای شرکت‌کنندگان حاوی اشاراتی در مورد رفتار موردانتظار محقق باشد، و شرکت‌کنندگان ممکن است بخواهند (با انگیزه‌های انساندوستی یا حس اطاعت) خود را با انتظارت محقق همخوان نمایند. رفتار واکنشی اثر دارونما را نیز توضیح می‌دهد، زیرا اثر دارونما به دلیل محتویات فعال درمان نیست (Shapiro & shapiro, 1997; L. White, Tursky, & Schwartz, 1985). در تحقیقات دارویی خود *عمل* دارو دادن می‌تواند باعث بهبود حال بیمار شود، حتی اگر داروی موردنظر تنها حاوی کمی شکر باشد (یک دارونما). مطالعات روزنبرگ (Rosenberg, 1969) شواهدی ارایه می‌کند که نشان می‌دهد پاسخ‌دهندگان نگران ارزیابی فرد آزمون‌گیرنده یا کارشناس، از پاسخهایشان هستند، و ممکن است بخواهند به گونه‌ای پاسخهایشان را تنظیم کنند که نشان‌دهنده کارآمدی و سلامت روان آنها باشد.

روزنتال و روزنو (Rosenthal & Rosnow, 1991) راههای بسیاری را برای کاهش اینگونه مشکلات پیشنهاد می‌کنند، از جمله راه‌حلهایی که پیش از این برای تغییرات خود-ابرازی واکنشی ارائه شد. بعلاوه راههای دیگری نیز وجو دارد که در ادامه به آنها اشاره می‌کنیم: (۱) با اندازه‌گیری متغیر وابسته، جایی بیرون از مختصات آزمایش، کاری کنیم که متغیر وابسته کمتر قابل‌شناسایی یا حدس‌زدن از سوی پاسخ‌دهنده باشد؛ (۲) اندازه‌گیری نتایج با فاصله زمانی زیاد [بعد از انجام مداخله] انجام شود؛ (۳) از انجام پیش‌آزمونهایی که حاوی اشاراتی در مورد نتایج موردانتظار از آزمایش است، اجتناب کنیم؛ (۴) از طرح چهار عاملی سولومون برای ارزیابی وجود احتمالی مشکل استفاده شود؛ (۵) آزمون را استاندارد کرده، و تعامل آزمونگر با پاسخ‌دهندگان را کاهش دهیم؛ (۶) از فرایندها یا داستانهای پوششی که باعث می‌شود آزمونگران و پاسخ‌دهندگان نتوانند فرضیات مطالعه را حدس بزنند، استفاده کنیم؛ (۷) در مواردی که منع اخلاقی ندارد، با دادن اطلاعات غلط در مورد فرضیات

191 Anchoring
192 Reactivity

آزمایش، پاسخ‌دهندگان را فریب دهیم؛ (۸) از شرکت‌کنندگان شبه‌کنترل استفاده کنیم. به این شرکت‌کنندگان در مورد فرایند آزمایش توضیحاتی ارائه شده، و از آنها پرسیده می‌شود که فکر می‌کنند باید چطور پاسخ بدهند؛ (۹) راهی پیش‌آزمونی برای اغناء و اشباع کردن میل شرکت‌کنندگان برای خوشنود کردن آزمایشگر بیابیم؛ (۱۰) شرایط آزمون را امن و غیر تهدیدآمیز کنیم، به نحوی که نگرانی و نیاز به حدس زدن هدف ارزیابی را کاهش دهیم؛ این کار را می‌توانیم از طریق تضمین گمنامی و محرمانه بودن پاسخها انجام دهیم. این راه حلها در بهترین حالت، نسبی هستند و غیرممکن است بتوانیم کاری کنیم افراد حدسها و فرضیات خود را در مورد مداخلهها نداشته باشند؛ زیرا این کار در تحقیقات میدانی اغلب غیرممکن، غیراخلاقی و مغرضانه است.

انتظارات آزمونگر

روزنتال (Rosenthal, 1956) دسته مشابهی از مشکلات را مورد بحث قرار می‌دهد، و از آنها به عنوان انتظارات آزمونگر یاد می‌کند. انتظارات آزمونگر بخشی از بسته درمانی مولکولی بوده، و می‌تواند نتایج را تحت تأثیر قرار دهد. روزنتال اولین بار در روانشناسی کلینیکی و در زمان انجام مطالعات خود در زمینه برانگیختن آزمایشی مکانیزمهای دفاعی، متوجه این مسأله شد. او این ایده را به صورت گسترده‌ای در تحقیقات آزمایشگاهی، و علی‌الخصوص در روانشناسی اجتماعی بسط داد. اما این مسأله در تحقیقات میدانی نیز مشاهده شده است. برای مثال در آموزش، این مسأله خود را به صورت اثر پیگمالیون[۱۹۳] نشان می‌دهد. این اثر به این پدیده اشاره دارد که انتظارات معلمین نسبت به دستاوردهای علمی دانش‌آموزان تبدیل به (ارزیابی‌های) گمانهای هماهنگ با انتظارات در آنها می‌شود. آن بخش از اثرات دارونما که توسط آزمونگر ایجاد می‌شوند (مانند زمانی که پرستار به فرد می‌گوید قرص موردنظر کمک‌کننده خواهد بود، در حالی که قرص موردنظر یک دارونمای بی‌اثر بوده است)، نیز در زمره این دسته از سوگیریها قرار می‌گیرند. برای مقابله با اثر این مشکل، و یا حداقل کاهش اثرات آن، روزنتال و روسنو (Rosenthal & Rosnow, 1991) پنج راه را پیشنهاد می‌کنند: (۱) استفاده از تعداد بیشتری آزمونگر، علی‌الخصوص اگر انتظارات آنها قابل‌دستکاری یا مطالعه‌کردن است؛ (۲) تحت‌نظر قرار دادن آزمونگر برای تشخیص و کاهش رفتارهای القاءکننده انتظارات؛ (۳) استفاده از رویه‌های پوششی که در آن آزمونگرها و مجریان درمان فرضیات مطالعه را نمی‌دانند؛ (۴) به حداقل رساندن تعامل و تماس میان آزمونگر و پاسخ‌دهندگان؛ استفاده از گروههای کنترل مانند کنترل دارونما برای ارزیابی احتمال وجود این مشکلات.

اثرات بدیع بودن[۱۹۴] و اختلال

193 Pygmalion effect
194 Novelty

براخت و گلس (Bracht & Glass, 1968) بر این باورند که در هنگام معرفی یک نوآوری جدید، موجی از انرژی، هیجان و اشتیاق در میان محققین برانگیخته می‌شود، که این خود منجر به موفقیت خواهد شد؛ علی‌الخصوص اگر تا پیش از این، نوآوری چندانی در آن زمینه رخ نداده باشد.[۱۹۵] اما بعد از سالیان متمادی نوآروی، ممکن است معرفی یک نوآوری جدید دیگر واکنشهای مثبت چندانی در پی نداشته باشد، و این باعث کاسته شدن از اثربخشی درمان موردنظر شود. در نقطه مقابل، معرفی یک نوآوری می‌تواند کاملاً مخرب باشد، خصوصاً اگر مانع بکارگیری خدمات اثربخش رایج فعلی باشد. در نتیجه، در اینجا نوآوری کمتر اثربخش خواهد بود. بدیع بودن و اختلال هر دو بخشی از بسته مولکولی مداخله هستند.

همسان‌سازی جبرانی

زمانی که در جریان اجرای یک مداخله یا درمان، خدمات یا محصولات مطلوبی را به طور نابرابر در اختیار گروهی قرار می‌دهیم، این امر می‌تواند موجبات برانگیختن مقاومت در میان اجراکنندگان تحقیق و یا کارکنان شود (Steven, 1994).[۱۹۶] برای مثال شومان و همکارانش (Schumacher et al., 1994) به مطالعه‌ای اشاره می‌کنند که در آن، اثر خدمات روزانه معمولی ارائه شده به افراد بیخانمان معتاد با شرایط خدمات سطح بالاتر روزانه مقایسه می‌شد. ارائه‌کنندگان خدمات درباره نابرابری بوجود آمده در ارائه خدمات اعتراض کرده، و خود شروع به ارائه خدمات سطح بالاتر به گروه دریافت‌کننده خدمات عادی کردند. در نتیجه تفاوت برنامه‌ریزی شده در آزمایش از میان رفت. همسان‌سازی می‌تواند به شکل دریغ کردن خدمات سطح بالاتر از گروه آزمون نیز بروز پیدا کند [یعنی عکس حالت قبلی]. در مطالعه‌ای، وکلای شاغل در یک دفتر منطقه‌ای حقوقی بر این باور بودند که شرایط آزمون بیش از اندازه به نفع موکلانشان است، و در نتیجه، از انجام توافق قبل از محاکمه[۱۹۷] برای آنها امتناع می‌کردند (Wallace, 1987). اینگونه نابرابریهای متمرکز می‌تواند دلیل اکراه برخی مجریان برای بکارگیری

---

۱۹۵ نمونه ای معروف از این تهدید اثر هاثورن نام دارد. در تفسیرهای اولیه آنچه در آزمایشی که در مرکز هاثورن شرکت وسترن الکتریک رخ داد، محققین چنین نتیجه گیری کردند که شرکت کنندگان (صرفنظر از هرگونه مداخله اعمال شده بر روی آنها) به دلیل توجه ویژه ای که به آنها شده بود کارایی خود را افزایش دادند. این تفسیر بعدها توسط دیگر محققین به چالش کشده شد اما نام «اثر هاثورن» همچنان برای این پدیده بکار گرفته می شود.

۱۹۶ بحث قبلی کوک و کمپبل (۱۹۷۹) در مورد این تهدید و سه تهدید بعدی می تواند برای خوانندگان گمراه کننده باشد. ممکن است فکر کنند که این تهدیدات تنها برای مطالعات با تخصیص تصادفی اتفاق می افتد. بلکه بالعکس این تهدید می تواند برای هر مطالعه ای که در آن شرکت کنندگان نسبت به تبعیضهای میان افراد از نظر دریافت انواع مختلف خدمات آگاه هستند، رخ دهد. این مقایسه ها در مطالعات شبه–آزمایشی نیز رخ می دهد و حتی مختص مطالعات پژوهشی نیز نیست (به عنوان نمونه نگاه کنید به شاپیرو (۱۹۸۴) درمورد طرح های رگرسیون ناپیوستگی).

۱۹۷ توضیح مترجم: توافق قبل از محاکمه (که طی آن متهم به گناه کوچکتری اقرار می‌کند و در عوض دادستان نیز از اتهامات شدیدتر او میگذرد و در نتیجه محاکمه زودتر تمام می‌شود)

تخصیص تصادفی باشد، در مواقعی که کارکنان بر این باورند که مراجعان آنها درمانی را بیش از درمان دیگر می‌خواهند. مصاحبه با مجریان و کارکنان راهی بسیار ارزشمند برای ارزیابی احتمال وجود این مشکل است.

رقابت جبرانی

تخصیص عمومی افراد به شرایط آزمون و کنترل می‌تواند موجب شکل‌گیری نوعی رقابت اجتماعی شود؛ بطوریکه گروه کنترل بخواهد نشان دهد که علی‌رغم دریافت نکردن مزایای مداخله، همچنان قادر است به خوبی گروه آزمایش عمل کند. سارتسکای (Saretsky, 1972) این اثر را «اثر جان هنری» می‌نامد. عنوان این اثر از نام آهنگری گرفته شد که زمانی که دریافت که نتایج کارش با دریل بخار مقایسه می‌شود، آنقدر سخت کار کرد که عملکردش از دریل بخار به مراتب بهتر بود، و در نهایت از کار زیاد درگذشت. سارتسکای همچنین به نتایج آزمایشی آموزشی اشاره می‌کند که در آن موفقیت عملکرد معلمین تجاری گروه آزمون (معلمین تجاری در ازای میزان یادگیری اتفاق افتاده در دانش آموزان دستمزد دریافت می‌کنند)، امنیت شغلی معلمین گروه کنترل را تهدید می‌کرد؛ معلمینی که ممکن بود با معلمین تجاری جایگزین شوند. در این آزمایش، معلمین گروه کنترل برای جلوگیری از احتمال از دست دادن کارشان با راندمانی بسیار بهتر از معمول کار می‌کردند. سارتسکای (Saretsky, 1972)، فترمن (Fetterman, 1982) و والتر و روس (Walter & Ross, 1982) به مثالهای دیگری نیز اشاره می‌کنند. روشهای کیفی از قبیل مصاحبه‌های باز و مشاهده مستقیم می‌تواند به کشف وجود چنین اثراتی کمک نماید. سارتسکای (Saretsky, 1972) تلاش کرد تا احتمال وجود این اثر را از طریق مقایسه عملکرد فعلی گروه کنترل با عملکرد همان گروه در سالهای قبل از شروع آزمون بررسی نماید.

تضعیف روحیه ناشی از رنجش و ناراحتی

بر خلاف مورد قبلی، افراد گروهی که مداخله با سطح مطلوبیت پایینتر را دریافت می‌کنند، و یا اصلاً مداخله را دریافت نمی‌کنند ممکن است دچار رنجش خاطر و تضعیف‌روحیه شوند، و در نتیجه، پاسخهای خود به نتایج را تغییر دهند. مطالعه فترمن (Fetterman, 1982) به نتایج ارزیابی یک برنامه آزمایشی آموزشی می‌پردازد. در این برنامه به افرادی که در دروه دبیرستان ترک تحصیل کرده بودند، در جریان جهت‌یابی شغلی این شانس دوباره داده می‌شد که تحصیلات دبیرستان خود را به پایان رسانده و دیپلم بگیرند. اگرچه طراحی آزمایش به گونه‌ای بود که تنها یک چهارم از افراد به گروه کنترل تخصیص داده شوند تا میزان مشارکت حداکثر شود، اما آنها که به گروه کنترل تخصیص داده شده بودند، عمیقاً رنجیده خاطر شدند. بسیاری از آنها اعتماد به نفس تحصیلی پایینی داشتند، و می‌باید تمام شجاعتشان را جمع می‌کردند تا بتوانند از این شانس مجدد استفاده کنند، و شاید این نه دومین شانس این افراد، بلکه آخرین شانس آنها برای به اتمام رساندن تحصیلات متوسطه به حساب می‌آمد. تضعیف روحیه ناشی از رنجش همواره به این میزان جدی نیست، اما این مثالها می‌تواند مسائل اخلاقی

همراه با آزمایشها را نشان دهد. یقیناً واکنش همه افراد به یک شکل نخواهد بود. مطالعه لم، هارتول و جکل (Lam, Hartwell & Jekel, 1994) نشان داد که افرادی که از ادامه مطالعه امتناع کرده بودند، واکنشهای متفاوتی داشتند. در نهایت، نتایج مطالعه شوماخر و همکارانش (Schumacher et al., 1994) نیز نشان می‌دهد که تضعیف روحیه ناشی از رنجش در مواردی که مداخله مثبت و مطلوب است، بیشتر احتمال وقوع دارد. یافته‌های این مطالعه حاکی از آن است که انتظارات مراجعین برای خدمات بهتر بالا می‌رفت، اما بعد با کاهش بودجه‌ها به یکباره کاهش پیدا می‌کرد، و مقاومت افراد نسبت به امکانات اقامتی تدارک دیده‌شده افزایش می‌یافت. مشکلات واکنشی نه تنها در واکنش به دیگر گروهها، بلکه در قیاس با آرزوهای خود فرد برای آینده نیز ایجاد می‌شود.

انتشار مداخله

برخی مواقع شرکت‌کنندگان در یک گروه (شرایط) آزمایشی، بعضی یا تمام مداخله دیگر شرایط را دریافت می‌کنند. برای مثال، در برنامه رفاه تجاری فلوریدا برای آزمایش کاری، نزدیک یک چهارم تمام شرکت‌کنندگان گروه کنترل به صورت متقاطع[۱۹۸] مداخله آموزش ضمن‌خدمت دریافت کردند (D. Greenberg & Shroder, 1997). اگرچه محققین در این مطالعه توانستند این افراد را شناسایی کنند، اما این افراد معمولاً از ترس اینکه محققین جلوی انتشار مداخله را نگیرند، این کار را به طور مخفیانه انجام می‌دهند؛ در نتیجه محققین معمولاً نمی‌توانند آنها را شناسایی کنند. این مشکل زمانی حادتر می‌شود که گروههای کنترل و آزمون از نظر مکانی به یکدیگر نزدیک باشند، و یا اینکه امکان تعامل و ارتباط با یکدیگر را دارند. برای مثال، فرض کنید از [ساکنان] ماساچوست به عنوان گروه کنترل برای بررسی اثر قانون سقط جنین در نیویورک استفاده شود، اگر افراد بتوانند براحتی برای انجام سقط جنین از ماساچوست به نیویورک مسافرت نمایند، اثرات حقیقی قانون مبهم خواهد بود. همچنین خطر انتشار در مواقعی که مجریان آزمایش برای گروه آزمون و کنترل یکسان هستند، امکان بروز بیشتری دارد. مانند موردمطالعه‌ای که رفتاردرمانی را با رواندرمانی الکتریکی مقایسه می‌کرد. درمانگر یکسانی دو نوع درمان را ارائه می‌کرد، و در جریان ارایه مداخله رواندرمانی الکتریکی به میزان زیادی از تکنیکهای رواندرمانی رفتاری استفاده می‌کرد (Kazdin, 1992). بهترین راه برای جلوگیری از انتشار، به حداقل رساندن اثرات مشترک در میان شرایط مختلف آزمایش (مثلاً از طریق استفاده از درمانگران متفاوت)، و جداکردن شرکت‌کنندگان در هر یک از گروهها از افراد دیگر گروهها (با استفاده از مکانهای جغرافیایی مجزا) است. اگر انجام این کار امکان‌پذیر نباشد، اندازه‌گیری اجرای مداخله در دو گروه می‌تواند مفید باشد. به این صورت که عدم وجود تفاوت یا تفاوت آزمایشی اندک در مقیاسهای اجرا، دال بر این است که انتشار رخ داده است (نگاه کنید به فصل ۱۰).

198 cross over

روایی سازه، تعدیل[۱۹۹] **پیش‌آزمایشی، و تعیین**[۲۰۰] **پس‌آزمایشی**

فرایند ارزیابی و درک سازه‌ها هیچگاه پایان نمی‌یابد. راه حلهای ارتقاء روایی سازه تا به حال بر این تأکید داشته‌اند که قبل از شروع آزمایش محقق باید با نگاهی نقادانه (۱) به اینکه سازه‌ها چطور باید تعریف شوند، فکر کند؛ (۲) سازه‌های مطالعه را از سازه‌های مرتبط یا مشابه تمیز دهد؛ و (۳) تصمیم بگیرد که چطور باید هر کدام از سازه‌های موردنظر را نمایه‌گذاری[۲۰۱] کند. این سه کار را می‌توان [تعیین] دامنه کاربرد موردنظر [سازه] نامید. علاوه بر این موارد، در این کتاب بر دو مورد بیشتر نیز تأکید می‌شود، (۴) نیاز به بکارگیری عملیاتهای متعدد برای نمایه‌کردن هر سازه، البته زمانی که امکانپذیر باشد (مثلاً مقیاسها، افراد، مداخلهها و مختصات متعدد) و همچنین زمانی که هیچکدام از گزینه‌های عملیاتی به وضوح بهتر از مابقی گزینه‌ها نیست؛ و (۵) نیاز به اطمینان حاصل کردن از اینکه هر کدام از این عملیات متعدد منعکس‌کننده روشهای متعدد باشد، به گونه‌ای که متغییرهای مزاحم ناشی از بکارگیری یک روش (مثلاً سوگیری ناشی از خود-اظهاری) را بتوان به نحو بهتری ارزیابی و شناسایی کرد.

پس از آنکه داده‌ها جمع‌آوری و به دقت تحلیل شدند، محققین می‌توانند بررسی کنند تا چه اندازه سازه موردنظر (به صورتی که از ابتدا مفهوم‌سازی شده بود) محقق شده است (دامنه کاربرد بدست آمده)؛ و در صورت لزوم آن را موردبازنگری قرار دهند. لزوم انجام این بازنگری از آنجاست که احتمال دارد عملیات برنامه‌ریزی شده به صورتی که از ابتدا برنامه‌ریزی شده بود پیش نرفته باشد، و یا اینکه شواهد حاکی از آن باشد که مطالعه در واقع سازه‌هایی غیر از سازه موردنظر آزمایش را انعکاس می‌دهد. بنابراین، بازتعریف سازه پس از اتمام آزمایش اجتناب‌ناپذیر به نظر می‌رسد؛ علی‌الخصوص در برنامه‌های مطالعاتی[۲۰۲] [در مقایسه با یک مطالعه منفرد]. فرض کنید در آزمایشی که با هدف مقایسه تعامل‌کنندگان با سطح اعتبار بالا و پایین انجام می‌شود، تفاوت معناداری در متغیر نتیجه‌ای مشاهده شود. اگر یک مقیاس پایای اندازه‌گیری سطح اعتبار تعامل کننده‌گان نشان دهد که ادراک شرکت‌کنندگان نسبت به سطح اعتبار تعامل‌کننده در شرایط مختلف آزمایش (بالا یا پایین) تفاوتی با یکدیگر ندارد، محقق باید با هر ابزاری که در اختیار دارد، این سوال را بررسی کند که اگر اعتبار نقشی نداشته است، پس اثر مشاهده‌شده در متغیر نتیجه‌ای در اثر چه عاملی بوده است؟ یا فرض کنیم که یک مداخله روی دو نسخه محاسبه شده (با پایایی) یک سازه اثر دارد، اما سه ویرایش دیگر از آن سازه را تحت تاثیر قرار نمی‌دهد. آزمایش فلدمن (Feldman, 1986) در بوستون، آتن و پاریس، پنج مقیاس متفاوت را برای اندازه‌گیری سازه میزان همکاری (بر مبنای درکی که در ابتدای مطالعه از سازه وجود داشت) مورد استفاده قرار داد تا ببیند آیا

199 Tailoring
200 Specification
201 Index
202 Programs of research

افراد یک کشور (هموطنان) میزان بیشتری از همکاری را در قیاس با خارجیها دریافت می‌کنند؟ این پنج مقیاس عبارت بودند از نشان دادن نشانی، لطف کردن به فردی از طریق پست کردن یک نامه گم‌شده، پس دادن بقیه پول به درستی وقتی که فرد ملزم به انجام این کار نیست، پس دادن پولی که فرد می‌تواند به سادگی و از روی اشتباه ادعا کند که پول خودش بوده و گرفتن مقدار کرایه درست از مسافران. نتایج تحلیل داده‌ها نشان داد که دو سنجه «نشان دادن نشانی» و «لطف کردن به فردی از طریق پست کردن یک نامه گم‌شده»، در مقایسه با سه مقیاس دیگر، ارتباط متفاوتی با مداخله آزمایشی دارند. این مسأله فدلمن را واردار کرد تا دو سازه همکاری تعریف کند؛ یکی اقدامات کم‌هزینه و دیگری نادیده گرفتن منافع مالی خود). اگرچه، فرایند ساختن فرضیات مرتبط با سازه ها[۲۰۳] و آزمون اینکه عملیات آزمایش تا چه اندازه متناسب با این سازهاست، چه قبل از آغاز مطالعه و چه بعد از اخذ داده‌ها، مشابه است.

همزمان با پایان مطالعه، لاجرم اختلافاتی بر سر اینکه مطالعه موردنظر تا چه اندازه توانسته است انعکاس‌دهنده سازه موردنظر باشد، بروز می‌کند. نقدهایی وارده معمولاً مطالعه را به این متهم می‌کنند که سازه‌هایی متفاوت از سازه موردنظر مطالعه را مفهوم‌سازی و عملیاتی کرده است. از آنجا که روایی سازه نیازمند آن است که معنای عملیات تحقیق به طور اجتماعی ساخته و بازساخته شود، یافتن تصمیمات یا راه‌حل‌هایی که ماندگار بوده، و برای مدت‌زمان طولانی قابل‌استفاده باشند، ندرتاً امکانپذیر است؛ و سازه‌ها عموماً نیازمند بازنگری هستند.

خوشبختانه، این اختلاف‌نظرها در مورد ترکیب سازه‌ها و یا بهترین راه برای اندازه‌گیری آنها، زمینه را برای انجام استنباطهای بهتر در مورد سازه فراهم می‌کند، چون نه تنها می‌توان این استنباطها را در طول نمودهای عملیاتی مختلف (و دارای همپوشانی) یک تعریف آزمون کرد، بلکه می‌توان آنها را در میان تعاریف متفاوت (اما دارای همپوشانی) همان سازه نیز بررسی کرد. برای مثال، جوامع زبانی متعددی با این موضوع که تمایل به آسیب رساندن را به عنوان بخشی از سازه پرخاشگری بدانیم، موافق نیستند. تنها زمانی می‌توانیم بدون نگرانی چنین تمایلی را از تعریف سازه پرخاشگری حذف کنیم که به طور عینی دریافته باشیم این تمایل تأثیر ناچیزی در نتایج مطالعه دارد. بنابراین، اختلاف‌نظرها درباره تعاریف سازه‌ها بطور بالقوه بسیار ارزشمند هستند.

روایی بیرونی

روایی بیرونی عبارتست از اینکه تا چه میزان یک رابطه علّی برای افراد، مختصات، مداخله‌ها و نتایج مختلف صدق می‌کند. برای مثال، آزمایش پایش آموزش ضمن خدمت شاغلین فصلی، افراد بالغ ۱۸ تا ۴۰ ساله دچار عقب‌ماندگی ذهنی را به طور تصادفی به دو گروه تخصیص می‌داد: گروه کنترل که در آن افراد خدمات معمولی دریافت می‌کردند و گروه آزمون که در آن افراد آموزش ضمن خدمت را همراه با کار دائمی دریافت می‌کردند

203 Hypothesizing constructs

(Greenberg & Shroder, 1997). نتایج نشان داد که مداخله موردنظر، جایابی شغلی و درآمد افراد را ارتقاء داد. اما محققین ملاحضات فراوانی را در مورد روایی بیرونی اثر مشاهده‌شده مطرح کردند. برای مثال، نتایج آنها نشان می‌داد که اثرات برای شرکت‌کنندگان با IQ بالاتر قابل‌توجه‌تر، و برای افرادی که IQ آنها کمتر از ۴۰ بود، بسیار ناچیز و یا صفر بود. ضمناً نتایج مقایسه‌های بین مکانی نشان می‌داد که نرخ موفقیت مشاهده‌شده تا حد زیادی وابسته به نوع مکان جایابی شغلی شرکت‌کنندگان است. محققین همچنین سؤالات دیگری در خصوص روایی بیرونی یافته‌های خود مطرح کردند، که داده‌های موجود اجازه بررسی آنها را نمی‌داد. برای مثال، برنامه در ۱۲ جا در آمریکا اجرا شده بود، اما هیچکدام از این مکانها در جنوب آمریکا نبود. بعلاوه، تنها ۵ درصد از کسانی که دعوتنامه دریافت کرده بودند، برای شرکت در آزمایش داوطلب شده بودند، و از آن میان، دو سوم افراد بدلیل واجد شرایط نبودن (از جمله شرایط مذکور، نداشتن مشکلات احساسی جدی و احتمال سود بردن از مداخله) از مطالعه کنار گذاشته شده بودند. اینکه آیا نتایج آزمایش برای افراد با آسیبهای احساسی جدیتر و افراد غیرداوطلب همچنان برقرار است، محل سؤال بود. علاوه بر آن، محققین گزارش کردند که افراد داوطلب ماجراجوتر بوده، و تمایل داشتند تا از محلهای اشتغال تحت حمایت سنتی به محیطهای واقعی شغلی منتقل شوند. محققین این سوال را مطرح می‌کنند که آیا منافع مشاهده‌شده شامل افراد با سطح پایین‌تر از ماجراجویی نیز خواهد بود؟

همانطور که از این مثال برمی‌آید، سؤالات روایی بیرونی به این می‌پردازند که آیا رابطه علّی تحت بررسی، (۱) برای سطوح مختلف[۲۰۴] افراد، مختصات، مداخلهها، و نتایج لحاظ شده در آزمایش؛ و (۲) برای افراد، مختصات، مداخلهها، و نتایج لحاظ نشده در آزمایش برقرار است؟ هدف تعمیم‌ها می‌تواند بسیار متنوع باشد:

- *از محدود به وسیع:* به عنوان نمونه از افراد، مختصات، مداخلهها، و نتایج یک آزمایش به جمعیت بزرگتر؛ مانند زمانی که سیاستگذاران کنجکاو هستند بدانند آیا نتایج بدست‌آمده از آزمایشهای حفظ درآمد در نیوجرسی، سیاتل و دنور را می‌توان به کل جمعیت آمریک تعمیم داد؟
- *از وسیع به محدود:* از نمونه آزمایشی به گروهی کوچکتر، و یا حتی به یک فرد؛ مانند زمانی که یک فرد مبتلا به سرطان سینه پیشرفته می‌پرسد آیا درمان تازه کشف‌شده که باعث افزایش طول عمر بیماران در دیگر نقاط دنیا شده برای افزایش طول عمر وی نیز- با در نظر گرفتن پاتولوژی، سطح پیشرفت بیماری و درمانهای بکارگرفته‌شده قبلی- سودمند است؟
- *در سطح مشابه:* از یک نمونه آزمایشی به نمونه آزمایشی دیگر با سطح مشابهی از مشخصات؛ مثل وقتیکه استاندار ایالتی مایل است رفرم های انجام‌شده در نظام رفاه در یک ایالت شبیه به ایالت خود (از نظر اندازه) را بکار بگیرد. رفرمهایی که بر اساس یافته‌های آزمایشی تأییدکننده آن رفرم [در ایالت مورد الگوبرداری] انجام شده است.

204 Variation

— *به یک نوع مشابه یا متفاوت:* در سه مورد بالا اهداف تعمیم ممکن است به نمونه‌های آزمایشی شبیه باشند (مانند مرد متقاضی شغل در سیاتل به مرد متقاضی شغل در آمریکا)، یا خیلی متفاوت باشند (از مردان آفریقایی-آمریکایی در نیوجرسی به زنان اسپانیای-آمریکایی در هیوستون).

— *نمونه تصادفی به اعضای جامعه:* در موارد نادری که نمونه‌گیری تصادفی در آنها انجام شده، می‌توان نتایج بدست‌آمده از نمونه تصادفی را به اعضای جامعه‌ای که نمونه از آن گرفته‌شده تعمیم داد

کرونباخ و همکارانش (Cronbach et al., 1980; Cronbach, 1982) بر این باورند که اغلب سؤالات روایی بیرونی در مورد افراد، مختصات، مداخلهها، و نتایجی مطرح می‌شود که در آزمایش موردمطالعه قرار نگرفته‌اند. زیرا این سؤالات معمولاً هنگامی مطرح می‌شوند که آزمایش پایان‌یافته است، و برای آنکه در مطالعه لحاظ شوند بسیار دیر شده است. برخی دانشمندان طرح این نوع از سؤالات روایی بیرونی را رد می‌کنند (به جز در مواردی که نمونه‌گیری تصادفی انجام شده باشد). این افراد چنین استدلال می‌کنند که محققین تنها در مقابل سؤالاتی که خودشان مطرح می‌کنند، و در جریان مطالعه خود موردبررسی قرار می‌دهند، مسئول خواهند بود، و در قبال سؤالاتی که بعدها دیگران در مورد شرایط کاربردهایی متفاوت از کاربردهای اولیه نتایج مطرح می‌کنند، مسئولیتی ندارند. به زعم این دانشمندان، استنباط در مورد کاربردهای هنوز-کشف-نشده وظیفه علم نیست، مگر زمانی که بتوان از طریق بازتحلیل داده‌های موجود، و یا انجام مطالعات جدید برای این استنباطها، پاسخی پیدا کرد.

نگارندگان این کتاب با کرونباخ و همکارانش هم عقیده هستند. استنباط از روی مطالعات انجام‌شده به کاربردهای هنوز مطالعه‌نشده هم برای علم، و هم برای جامعه، ضروریست. برای مثال، در طول دو دهه اخیر در قرن بیستم، محققین بخش ارزیابی برنامه و روش‌شناسی در اداره حسابداری کل آمریکا مکرراً بر اساس مرور مطالعات انجام شده در گذشته، راهنماییها و پیشنهاداتی را در مورد سیاستهای مختلف به کنگره ارائه داده‌اند. مطالعاتی که غالباً همپوشانی‌هایی بخشی و جزئی با کاربردهای مدنظر کنگره داشتند (Droitcour, 1997). در حقیقت، اگر با نگاهی واقعگرایانه بنگریم، گوهر وجودی علم خلّاق آن است که با بسط فزاینده نظریه و آزمایش به قلمروهای آزمون‌نشده‌ای که -به نظر دانشمندان و با در نظر گرفتن دانش موجود- احتمال دارد ره‌آورد ارزشمندی داشته باشند، برنامه‌های تحقیقاتی را پیش برده و توسعه دهد (McGuire, 1997). اینگونه برونیابیها[۲۰۵] عموماً توجیه‌پذیر هستند، چون آنها واریانه‌های فزاینده در برخی، و نه تمامی مؤلفه‌های مطالعه هستند؛ چیزی که باعث می‌شود تا انجام این برونیابیها به چیزهایی که تا به حال مطالعه نشده‌اند توجیه‌پذیرتر و موجّه‌تر باشد. برای مثال، ممکن است این سؤال مطرح باشد که اثرات برنامه ممنوعیت استعمال دخانیات در محیطهای کاری

205 Extrapolation

که در شرکتهای خصوصی مورد بررسی قرار گرفته، قابل‌تعمیم به محیطهای دولتی نیز هست؟ حتی اگر محیطهای کاری دولتی هیچگاه پیش از این مورد مطالعه قرار نگرفته باشند، همچنان محیط کاری هستند، و احتمال دارد درمان و مشاهدات تا حد زیادی مشابه باشند، و افراد مطالعه‌شده از نظر بسیاری از مشخصه‌ها مانند سیگاری بودن یکسان باشند. سؤالات مرتبط با روایی بیرونی می‌تواند به این صورت نیز پرسیده شود که اگر تمامی مشخصه‌های یک مطالعه متفاوت بود، چه اتفاقی می‌افتاد؟ اما این شکل از سؤال روایی بیرونی آنقدر کمیاب است که حتی مثال خوبی از آن برای ذکر در اینجا نیافتیم.

از سوی دیگر، اینکه روایی بیرونی را تنها به سؤالات مرتبط با مواردی که هنوز مطالعه نشده‌اند محدود کنیم، نیز درست نیست. کمپبل و استنلی (Stanley & Campbell, 1963) در فرمول اولیه خود از روایی بیرونی به این سؤال که «این اثر به چه جمعیتها، مختصات، متغیرهای مداخله، و متغیرهای اندازه‌گیری قابل تعمیم است» اشاره‌ای نکرده‌اند. یقیناً، یکی از اهداف نظریه آنها اشاره به این بود که «راههای بسیاری وجود دارد که از آن طریق می‌توان آزمایشها را از نظر روایی بیرونی ارتقاء داد (ص ۱۷)». برای مثال، آنها ادعا می‌کنند روایی بیرونی مطالعات تکی در صورتی که «آزمایش اولیه پدیده را در طیف وسیعی از شرایط نشان داده باشد»، و همچنین در صورتیکه «شباهت حداکثری میان آزمایش و شرایط کاربرد نتایج وجود داشته باشد»، افزایش می‌یابد. این هدف که آزمایشها را به گونه‌ای طراحی کنیم که روایی بیرونی بالاتری داشته باشند، اساساً مفهوم جدید و بدیعی نیست. بالعکس، اغلب آزمایشها این مسأله را که آیا اثر مداخله در شرایط مختلف وجود دارد، را آزمون می‌کنند. بسیاری نیز این موضوع را که آیا نتایج بدست‌آمده برای افراد مختلف معنادار است، را گزارش می‌کنند- اگرچه تقسیم کردن نمونه توان [آماری] مطالعه را کاهش می‌دهد. آزمون پایداری اثر در حضور مداخلههای گوناگون، مختص مطالعاتی است که مداخلههای متعددی دارند، اما این مسأله به طور معمول در مطالعات علمی رخ می‌دهد (Wampold et al., 1997). همینطور آزمون اینکه روابط علّی تا چه اندازه در حضور مختصات متفاوت رخ می‌دهد نیز امری معمول است؛ مثلاً در زمینه آموزش (Raudenbush & Willms, 1995)، و یا در آزمایشهای بزرگ-مقیاس چند-مکانی درمانی و بهداشت عمومی (Ioannidis et al., 1999).

با این وجود، محدودیتهای روشنی برای بکارگیری این استراتژی وجود دارد. محققین معدودی تا آن اندازه همه چیزدان هستند که بتوانند تمامی شرایطی که احتمالاً یک رابطه علّی را تحت تاثیر قرار خواهد داد را پیش‌بینی کنند. حتی اگر تا به این اندازه همه‌چیزدان بودند نیز، یک راه حل کامل، آزمایش را ملزم به در برگرفتن طیف ناهمگونی از افراد، مداخلهها، مختصات، و متغییرهای نتیجه‌ای خواهد کرد. متنوع کردن نتایج و برون‌دادها معمولاً شدنی است. اما بکارگیری مکانهای مختلف، یا عملیاتی کردنهای مختلف از مداخله، و یا آزمون رابطه علّی با در نظر گرفتن مشخصات مختلف افراد تحت‌آزمون دشوار خواهد بود؛ و انجام همزمان تمامی این اقدامات همزمان با یگدیگر، به طور غیرقابل‌تحملی گران و از نظر لجیستیکی پیچیده می‌شود. حتی اگر آزمایشی

دربرگیرنده تنوع لازمه باشد، شناسایی اثرات متقابل دشوارتر از شناسایی اثرات اصلی مداخله خواهد بود. اگرچه با استفاده از برخی طرحهای آزمایشی می‌توان توان آزمایش را افزایش داد (West, Aiken, & Todd, 1993)، اما این طرحها باید پیش از آنکه مطالعه آغاز شود بکار گرفته شوند؛ مسأله‌ای که باعث می‌شود تا استفاده از این طرحها برای عمده سؤالات مرتبط با روایی بیرونی-که غالباً پس از پایان یک مطالعه مطرح می‌شوند- بی‌معنا باشد. علاوه بر این، محققین معمولاً دلیلی بسیارخوب برای متنوع نکردن تمامی مشخصات یک آزمایش دارند؛ و آن اینکه واریانس فرعی و بی‌ارتباط[۲۰۶] در مختصات و افراد موردآزمایش، خطری برای روایی نتایج آماری محسوب می‌شود. بنابراین، هنگامی که بدلیل انتظار وجود اثرات متقابل، نمونه‌گیریهای ناهمگون انجام می‌شود، اندازه نمونه باید به میزانی افزایش یابد که تضمین‌کننده توان آماری لازم باشد. این کار نیازمند پول فراوانی است که می‌توانست برای ارتقاء دیگر مشخصه‌های طرح آزمایش صرف شود. با توجه به محدودیت منابع، طراحی مطالعاتی که بتوانند سؤالات مرتبط با روایی بیرونی را افزایش دهند، اغلب در تضاد با دیگر اولیتهای طراحی که از اولویت بیشتری برخوردارند، قرار دارد.

برخی اوقات، زمانی که مطالعه اولیه متغیرهای مرتبط [با روایی بیرونی] را شامل می‌شود، اما آنها را تحلیل نکرده، و یا در گزارشات نیاورده است، محقق یا دیگر محققین می‌توانند در مطالعات ثانویه (Kiecolt & Nathan, 1990) داده‌ها را بازتحلیل کرده تا ببینند در صورت تغییر در متغیر، رابطه علّی چه تغییراتی می‌کند. برای مثال، مطالعه‌ای در مورد اثرات یک برنامه کاهش وزن خاص بر نمونه‌ای متشکل از مردان و زنان، ممکن است به این نتیجه منتهی شود که برنامه موردنظر اثر بخش بوده است. اگر پس از انجام مطالعه این سؤال بوجود بیاید که آیا نتایج برای زنان و مردان به طور مجزا همچنان برقرار است یا نه، می‌توان داده ها را مجدداً تحلیل کرده، و به این سوال پاسخ داد؛ البته به شرطی که داده‌ها در دسترس باشند، و به گونه‌ای کدگذاری شده باشند که این تفکیک میسر باشد.

البته معمولاً داده‌های اولیه در دسترس نبوده، و یا حاوی اطلاعات موردنیاز برای تحلیل‌های ثانویه نیستند. در چنین مواردی، مرور نتایج مطالعات منتشرشده در مورد آن سؤال خاص (منظور انجام متاآنالیز است)، منبع بسیار خوبی برای پاسخگوی به سؤالات روایی بیرونی خواهد بود. همانطور که کمپبل و استنلی (Campbell & Stanley, 1963) اشاره می کنند، «ما معمولاً تنها قسمت به قسمت، و در جریان آزمون و خطا می‌آموزیم که تا چه اندازه می‌توانیم داده‌های دارای روایی درونی را تعمیم دهیم (ص ۱۹)»؛ معمولاً در طول مطالعات متعددی که انواع متفاوتی از افراد، مختصات، مداخلهها و نتایج را در بر دارند. دانشمندان این کار را از طریق اجرای سری برنامه‌های پژوهشی [و نه پژوهش‌های منفرد] در طول دوران فعالیت حرفه‌ای خود انجام می‌دهند. فرایند زمانبری که بیشترین میزان کنترل را بر تعمیم خاص موردنظرشان فراهم می‌آورد. این کار را همچنین از طریق

206 Extraneous

ترکیب کارهای خود با پژوهشهای دیگر دانشمندان نیز انجام می‌دهند؛ ترکیب پژوهشهای بنیادی و کاربردی، یا مطالعات آزمایشگاهی و میدانی. دویر و فلشیجنی (Dwyer & Flesch-Janys, 1995) این کار را در جریان مرور اثرات عامل پرتغالی[۲۰۷] در ویتنام انجام دادند. و در نهایت، دانشمندان این کار را از طریق مرور کمّی آزمایشهای زیادی که به سوال مشترکی می‌پردازند انجام می‌دهند. اینگونه *متاآنالیزها* توجیه‌پذیرتر از انجام مطالعات ثانویه هستند؛ چون برای انجام آنها نیازی به داده‌های اولیه نخواهیم داشت. اگرچه متاآنالیز نیز مشکلات خود را دارد؛ کیفیت ضعیف گزارش‌دهی یا تحلیل آماری در برخی مطالعات، از جمله این مشکلات هستند. فصل ۱۳ تمامی موضوعات مرتبط با متاآنالیز را به تفصیل مورد بحث قرار خواهد داد.

عوامل تهدیدکننده روایی بیرونی

تخمین میزان تعمیم‌پذیربودن رابطه علّی به طیفهای متنوعی از افراد، مختصات، مداخلهها و متغیرهای نتیجه‌ای به لحاظ مفهومی، چیزی شبیه آزمون آماری برهم‌کنش[۲۰۸] است. اگر میان یک مداخله آموزشی، و طبقه اجتماعی کودکان برهم‌کنش وجود داشته باشد، نمی‌توان گفت که نتایج یکسانی در میان طبقات اجتماعی [مختلف] مشاهده خواهد شد. زیرا برهم‌کنش معنادار، اندازه اثرهای متفاوتی را برای طبقه‌های اجتماعی متفاوت نشان خواهند داد. در نتیجه، نگارندگان تهدیدات روایی بیرونی را در قالب برهم‌کنش روابط علّی با (۱) افراد، (۲) مختصات، (۳) مداخلهها و (۴) نتایج، بیان می‌کنند؛ که این شامل متغیرهای واسطه‌ای در آن روابط نیز می‌شود (نگاه کنید به جدول ۳.۲).

البته بکار بردن عبارت برهم‌کنش در نامگذاری این تهدیدها به معنی محدود کردن این تهدیدها به تعاملهای آماری نیست. بلکه این مفهوم در پس عبارت برهم‌کنش است که اهمیت دارد – یعنی جستجوی راههایی که از آن طریق یک رابطه علّی می‌تواند برای افراد، مختصات، مداخلهها، و نتایج مختلف تغییر کند یا نکند. اگر این سؤال را بتوان از طریق کمّی کردن رابطه برهم‌کنشی و آزمون آماری آن پاسخ داد، که بسیار خوب است، اما ناتوانی از انجام این کار نباید ما را از تلاش برای یافتن این تهدیدها بازدارد. برای مثال، در مورد تعمیم دادن به افراد، مختصات، مداخلهها، و نتایجی که مطالعه نشده‌اند، آزمون آماری برهم‌کنش امکانپذیر نیست. اما این موضوع مانع از آن نمی‌شود که محققین فرضیات موجّهی در مورد برهم‌کنشهای احتمالی بسازند. این فرضیات برخی اوقات بر مبنای تجربیات حرفه‌ای، و بعضی وقتها بر اساس مطالعات مرتبط انجام می‌شود، و با استفاده از آنها تعمیم نتایج تجربی موردنقد قرار گرفته، و ایده‌هایی برای طراحی مطالعات و آزمایشهای جدید شکل می‌گیرد. محقق نباید خود را برده معناداری آماری تعاملها بداند. روابط غیرمعنادار ممکن است حاکی از توان آماری پایین باشند. با همه این احوال، این نتایج ممکن است همچنان از نظر کاربردی اهمیت داشته باشند،

---

207 Agent Orange [توضیح مترجم: گیاهی سمی که در حملات شیمیایی آمریکا علیه ویتنام بکار گرفته شد.]

208 Interactions

چون می‌توانند زمینه‌ساز تحقیقات آتی باشند. در مقابل، تعاملهای معنادار نیز می‌توانند از نظر نظری و کاربرد بی‌اهمیت باشند. بنابراین، مسأله تنها معناداری آماری برهم‌کنشها نیست، بلکه مسأله اهمیت نظری و کاربردی آنهاست؛ نه اینکه فقط بتوان این روابط را در داده‌ها شناسایی کرد، بلکه اینکه بتوانیم از آنها برای ساختن مسیرهای پژوهشی جدید در رابطه با محدودیتهای روابط علّی بهره ببریم.

جدول ۳.۲: تهدیدات روایی بیرونی: چرا استنباطهای انجام‌شده در مورد برقراری نتایج مطالعه، برای افراد، مختصات، مداخلهها و نتایج مختلف، می‌تواند نادرست باشد؟

۱. *برهم‌کنش رابطه علّی با افراد:* اثری که در رابطه با برخی افراد خاص در یک مطالعه مشاهده می‌شود، ممکن است برای دیگر انواع افراد معنادار نباشد. یعنی اگر دیگر افراد مطالعه می‌شدند این اثر مشاهده نمی‌شد.
۲. *برهم‌کنش اثر علّی و انواع مداخله:* اثری که با برخی واریته‌های۲۰۹ یک مداخله بروز می‌کند، ممکن است با دیگر واریته‌های آن مداخله، و یا هنگام ترکیب آن مداخله با دیگر مداخلهها، و یا بکارگیری بخشی از مداخله (و نه تمام آن) اثربخش نباشد.
۳. *برهم‌کنش رابطه علّی با نتایج:* اثری که با در نظر گرفتن یک متغییر نتیجه‌ای خاص مشاهده شده است، ممکن است با در نظر گرفتن دیگر متغیرهای نتیجه‌ای برقرار نباشد.
۴. *برهم‌کنش رابطه علّی با مختصات آزمایش:* اثری که در یک نوع خاص از مختصات آزمایشی وجود دارد، ممکن است در دیگر انواع مختصات آزمایشی برقرار نباشد.
۵. *متغیرهای میانجی وابسته به زمینه۲۱۰:* متغیر میانجی در یک رابطه علّی در زمینه‌ای خاص، ممکن است در زمینه‌ای دیگر اثر میانجیگری نداشته باشد.

209 Variations
210 Context-dependent mediation

برهم‌کنش رابطه علّی با افراد

رابطه علّت و اثر موردنظر در مورد کدامیک از افراد صدق می‌کند؟ برای مثال، باور غالب در دهه هشتاد در آمریکا این بود که تحقیقات حوزه بهداشت و درمان به طور نامتوازنی بر روی مردان سفیدپوست انجام می‌شود (بطوری که به طنز گفته می‌شد حتی موشها هم سفید و نر هستند، چون اغلب موشهای مورد استفاده در آزمایشگاهها سفید بودند، و به منظور تأمین همگونی نمونه معمولاً از نوع نر آنها استفاده می‌شد). محققین روی این مسأله حساس شده بودند که مبادا نتایج بدست‌آمده از جامعه مردان سفید برای زنان و مردان دیگر نژادها برقرار نباشد. در نتیجه مؤسسه بهداشت و درمان ملّی آمریکا به طور رسمی اعلام کرد که این تنوع باید به طور سیتماتیک در مطالعات مورد بررسی قرار بگیرد (Hohman & Parron, 1996). حتی در آزمایشهایی که اختصاصاً شرکت‌کنندگان گروههای خاصی (مثلاً زنان سیاهپوست آفریقایی-آمریکایی) را مورد مطالعه قرار می‌داد نیز افراد مورد مطالعه ممکن بود در مقایسه با افرادی که مطالعه نمی‌شدند، به طور سیستماتیک متفاوت باشند. این افراد ممکن بود از جمله افراد داوطلب، افراد مالیخولیایی و خودبیمارانگار، خودنما، افرادی که از کارهای علمی عام-المنفعه لذت می‌برند، افرادی که به پولهایی که بابت انجام آزمایش داده می‌شود نیاز دارند، دانشجویانی که به نمره اضافی احتیاج دارند[۲۱۱]، افرادیکه به شدت نیاز به کمک دارند، و یا کسانی که هیچ کار دیگری برای انجام ندارند، باشند. مثلاً، در آزمایش برنامه کاری آرکانزاس، به طور برنامه‌ریزی‌شده افرادی برای مداخله انتخاب شدند که بیشترین میزان آمادگی برای اشتغال را داشتند.

اینطور خامه‌گیری[۲۱۲] از نمونه باعث می‌شود تا تخمینهای بدست آمده بالا باشد، به گونه‌ای که اگر افراد با آمادگی اشتغال کمتر انتخاب می‌شدند، این نتایج بدست نمی‌آمد. به همین ترتیب، هنگامی که واحد مورد آزمایش، یک مدرسه است، احتمالاً سازمانهای داوطلبی [در مقایسه با دیگر سازمانها]، برای شرکت در آزمون پیشتازتر و از خود مطمئن‌تر خواهند بود. برای مثال، کمپبل (Campbell, 1956) اگرچه با اداره پژوهشی ناوبری دریایی کار می‌کرد، اما نتوانست به ملوانان تخریب‌چی دسترسی داشته باشد، و مجبور شد از ملوانان زیردریایی با سطح وجدان و ارزشهای اخلاقی بسیار بالا استفاده کند. آیا می‌توان نتایج بدست‌آمده از چنین افرادی را به افراد با سطح ارزشهای اخلاقی و وجدان پایینتر تعمیم داد؟

برهم‌کنش رابطه علّی با طیف متنوعی از مداخلهها

اندازه و جهت رابطه علّی در حضور واریته‌های متفاوت مداخلههای مختلف تغییر می‌کند. برای مثال، کاهش اندازه کلاس زمانی اثربخش خواهد بود که در کنار سرمایه‌گذاری قابل‌توجه برای ساخت کلاسهای جدید، و

۲۱۱ – توضیح مترجم: در برخی دانشکده های علوم انسانی، شرکت در آزمایشهای در حال انجام توسط اعضای هیات علمی، جزء موظفی دانشجویان بوده و بخشی از نمره آنها محسوب می شود.

استخدام معلمین باتجربه قرار گیرد. اما اگر این سرمایه‌گذاری به درستی انجام نشود، مداخله کاهش اندازه کلاس (تعداد دانش‌آموزان) نیز اثربخش نخواهد بود. به همین ترتیب، کوتاه‌بودن دوره زمانی اغلب مداخلهها باعث می‌شود تا واکنش احتمالی افراد، متفاوت از حالتی باشد که در صورت طولانی بودن دوره زمانی مداخله رخ می‌داد. بنابراین، در آزمایش حفظ درآمد نیوجرسی، پاسخ‌دهندگان به درآمدی واکنش نشان دادند که برای سه سال تضمین شده بود. از آنجا که این تردید وجود داشت که پاسخ‌دهندگان در صورت طولانی‌تر بودن دوره مداخله، واکنش متفاوتی نشان بدهند، آزمایش بعدی که در سیاتل-دنور انجام شد خانواده‌هایی را مورد بررسی قرار داد که مزایای آنها برای بیست سال (تقریباً به طور مادام‌العمر) تضمین می‌شد. همچنین اثرات بدست آمده در آزمایش کوچک ممکن است بسیار متفاوت از اثراتی باشد که در جریان اجرای بزرگ مقیاس آن مداخله در جمعیت مشاهده خواهد شد. برای مثال، این اتفاق زمانی می‌تواند رخ دهد که یک مداخله اجتماعی تغییر نگرش و نُرمهای اجتماعی تنها در صورتی اثربخش باشد که در مقیاس وسیع اجرا شود. در چنین مواردی، آزمایشهای اجتماعی که در مقیاسی کوچکتر از مقیاس مدنظر دولت اجرا شده باشند، نمی‌توانند باعث تغییرات اجتماعی موردنظر باشند. در نهایت، این تهدید شامل برهم‌کنشهایی می‌شود که در هنگام بکارگیری چند مداخله بطور همزمان رخ می‌دهند. اثرات برهم‌کنشی داروها مثالی رایج از این دست اثرات هستند. یک دارو ممکن است به خودی خود اثری بسیار مثبت داشته باشد، اما هنگامی که در ترکیب با دیگر داروها مصرف می‌شود، می‌تواند مرگ‌آور بوده، و یا اثر آن به کلی از بین برود (مثل مصرف آنتی‌بیوتیکها با لبنیات). بالعکس، ترکیب داروهای درمان‌کننده ایدز می‌تواند تا حد زیادی میزان مرگ و میز ناشی از بیماری را کاهش دهد، اما هر کدام از داروها به تنهایی چندان اثربخش نیستند.

برهم‌کنش رابطه علّی با نتایج

آیا یک رابطه علّت و معلولی را می‌توان به نتایج (متغیرهای نتیجه‌ای) مختلف تعمیم داد؟ مثلاً در زمینه پژوهشهای سرطان، میزان اثربخشی درمان بسته به اینکه متغیر نتیجه را کیفیت زندگی در نظر بگیریم، یا پنج سال ماندگاری بدون متاستاز، و یا ماندگاری به طور کلی، متفاوت خواهد بود. اما یک فرد عادی تنها مورد آخر را به عنوان درمان شدن در نظر می‌گیرد. همینطور وقتی نتایج علوم اجتماعی به اطلاع مخاطبین می‌رسد، خیلی پیش می‌آید که نظراتی از این دست بشنویم که، «بله، من قبول دارم که برنامه آموزش اشتغال جوانان به آنها کمک می‌کند که بلافاصله بعد از اتمام تحصیلاتشان کار پیدا کنند، اما این چه ربطی به مهارتهای سازگاری شغلی مانند سروقت بودن، و توانایی پیروی از دستورات دارد؟» پاسخ به چنین سوالاتی تصویری کامل از کل اثر یک مداخله ارائه می‌کند. برخی اوقات اثر یک مداخله بر یک متغیر نتیجه‌ای مثبت، بر دیگری منفی، و بر سومی

212 Creaming

صفر است. به عنوان نمونه، در آزمایش حفظ درآمد اجراشده در نیوجرسی، حقوق پرداخت‌شده باعث شد تا ساعات کار بیرون از خانه زنان در این خانواده‌ها کاهش پیدا کند، و احتمال آنکه نوجوانان در این خانواده‌ها تحصیلات متوسطه خود را به اتمام برسانند، افزایش یابد؛ اما مداخله موردنظر هیچ اثری بر میزان مالکیت خانه و خرید لوازم خانگی اصلی در این خانواده‌ها نداشت. خوشبختانه متغیرهای نتیجه‌ای را آسانتر از دیگر اجزاء آزمایش می‌توان تغییر داد (Kreshaw & Fair, 1977; Watts & Rees, 1976). مشاوره با ذینفعان پیش از انجام مطالعه می‌تواند روشی بسیارخوب برای تضمین این باشد که سؤالات احتمالی مرتبط با تعمیم‌پذیری، برای انواع مختلف نتایج، در طراحی آزمایش پیش‌بینی شده است.

برهم‌کنش رابطه علّی با مختصات آزمایش

یک رابطه علّت و معلولی در چه مختصات آزمایشی بدست خواهد آمد؟ مثلاً کازدین (Kazdin, 1992) برنامه‌ای را روی معتادین مواد مخدر اجرا کرد. این برنامه در میان معتادین روستایی اثربخش بود، اما اثری بر معتادین شهری نداشت. شاید دلیل آن بود که مواد مخدر با سهولت بیشتری در محیطهای شهری در دسترس هستند. پاسخ به اینگونه سوالات را می‌توان از طریق اعمال تغییرات در مختصات آزمایش، و بررسی و تحلیل رابطه علّی در هر یک از آن مختصات [مختلف] بدست آورد. اما این کار اغلب هزینه‌بر است، بگونه‌ای که انجام این گزینه‌ها به ندرت توجیه‌پذیر هستند. با این وجود، بعضی اوقات یک مکان بزرگ (مانند دانشگاه)، زیرمجموعه‌هایی دارد (مانند دپارتمانهای مختلف) که از نظر عواملی که بر نتایج اثرگذار هستند، به طور طبیعی واریانس دارند. این محلها امکان برخی مطالعات تعمیم‌پذیری را میسر می‌سازند. مطالعات بزرگ چندمکانه[۲۱۳] نیز ظرفیت بررسی چنین مسائلی را داشته (Turpin & Sinacore, 1991)، و کار پیچیده و دشوار آشکار ساختن علّت وجود تفاوت میان مکانها را میسر می‌سازند (Raudenbush & Willms, 1991).

متغیرهای میانجی وابسته به زمینه

توضیح علّی به عنوان یکی از اصول پنجگانه تعمیم علّی در نظریه بنیادین، در فصل اول مورد اشاره قرار گرفت. با این وجود، این اصل را در فصل ۱۲، به تفصیل و با جزئیات مورد بحث قرار خواهیم داد. یکی از بخشهای توضیح علّی شناسایی فرایندهای میانجی است. ایده اصلی این است که مطالعات مرتبط با میانجیهای علّی، فرایندهای دخیل در شکل‌گیری و وقوع یک اثر را شناسایی می‌کنند. اگرچه، حتی اگر یک میانجی درست در یک زمینه خاص شناسایی شود، آن متغیر ممکن است در زمینه‌های دیگر اثر میانجی‌گری نداشته باشد. مثلاً نتایج مطالعه‌ای با موضوع بررسی اثرات یک برنامه بیمه‌درمانی جدید در بیمارستانهای خیریه ممکن است نشان دهد که هزینه‌ها در حال کاهش هستند. اگرچه این کاهش ممکن است به دلیل کاهش در میزان خدمات ارائه‌شده به

213 Multisite

بیماران اتفاق افتاده باشد. در این مثال، تغییرات زمینه‌ای در مختصات آزمایشها رخ داده، اما این اتفاق می‌تواند به معنی تغییر در شرکت‌کنندگان آزمایش، و یا در ماهیت مداخله و یا در متغیرهای نتیجه‌ای باشد.
وابستگی‌های زمینه‌ای در هرکدام از این موارد نیز اثر برهم‌کنشی محسوب می‌شود- در این مورد برهم‌کنش میان متغیر میانجی در رابطه علّی، با هرکدام از اجزاء زمینه آزمایش که دچار تغییر شده، روی داده است. اگر چنین متغیرهای میانجی‌ای را بتوان شناسایی کرد، و در زمینه‌های مختلف موردمطالعه قرار داد، ثبات رفتاری این متغیرها به عنوان یک میانجی را می‌توان با استفاده از مدلهای معادلات ساختاری چندگروهی مورد آزمون قرار داد.

ثبات اندازه اثر در مقابل ثبات جهت علّی

در این کتاب تهدیدات روایی بیرونی در قالب اثرهای برهم‌کنشی آورده شد. این اثرهای برهم‌کنشی باید تا چه اندازه بزرگ باشند تا تهدیدی برای تعمیم‌پذیری به شمار آیند؟ آیا تغییر اندکی در اندازه اثر را می‌توان به عنوان ناتوانی در تعمیم دادن قلمداد کرد؟ این سؤالات از نظر آماری اهمیت دارند، چون مطالعه‌ای با توان آماری بالا می‌تواند حتی تغییرات ناچیز در اندازه اثر را به ازای مقادیر مختلف یک متغیر تعدیلگر شناسایی کند. همچنین پاسخ به این سؤالات از نظر فلسفی نیز حائذ اهمیت است. چون براساس بسیاری از نظریه‌ها، جهان بدلیل ماهیتش پر است از برهم‌کنشها؛ به گونه‌ای که اثرات آماری اصلی ندرتاً می‌توانند دنیا را با درستی کامل تبیین کنند (Mackie, 1974). البته اهمیت این اثرها نسبی است، چون دانشمندانی مانند کرونباخ و اسنو (Cronbach & Snow, 1977) در زمینه آموزش، مگنوسن (Magnusson, 2000) در علوم رشد، و مک گوایر (McGuire, 1984) در روانشناسی اجتماعی، ادعا می‌کنند که وجود اثرهای برهم‌کنشی پیچیده نوعی قائده به حساب می‌آیند. پس اگر *ثبات اندازه اثر* را بعنوان معیار *استحکام*[۲۱۴] در نظر بگیریم، روابط علّی اندکی در دنیای اجتماعی قابل‌تعمیم خواهند بود.
با این وجود، نگارندگان این کتاب بر این باورند که تعمیم‌پذیری را می‌توان اغلب به صورت *ثبات جهت [رابطه] علّی* تعریف کرد. به این معنا که علامت (مثبت یا منفی) رابطه علّی به ازای مقادیر مختلف متغیر تعدیلگر، همچنان ثابت بماند. عوامل متعددی را می‌توان به عنوان پشتوانه برای این بحث مطرح کرد. اول اینکه بررسی علّی بسیاری از متاآنالیزها ما را متقاعد می‌کند که حداقل در موضوعاتی که در آنها گروههای مداخله و کنترل با یکدیگر مقایسه شده‌اند، علامت رابطه علّی در مطالعات منفرد متعدد ثابت مانده باشد، حتی زمانی که اندازه اثر به اندازه قابل‌توجهی تغییر کرده است (برای مثال، Shadish, 1992a). دوم، در دنیای سیاستگذاری اجتماعی، بسیار دشوار است که مقررات و قوانینی بسازیم که مناسب مقتضیات محلی باشد. در عوض، باید برنامه مشترکی

214 Robustness

در سطح کل کشور، یا ایالت به طور رسمی اشاعه و ترویج شود تا از نابرابریهای متمرکز میان مکان، افراد، و گروهها اجتناب شود. علیرغم واریانس اجتناب‌ناپذیر اندازه اثر از یک مکان به مکان دیگر، یا از یگ گروه از افراد به گروهی دیگر، و متغیرهای مختلف نتیجه‌ای و مداخلههای مختلف، سیاستگذاران همچنان به اثرات مطلوب امیدوارند. ترس آنها از این است که علامت رابطه علّی در شرایط متفاوت تغییر کند. سوم، نظریه‌های اصلی معمولاً حول آن نوع از روابط علّی شکل می‌گیرند که وقوع آنها قابل‌اتکاء [قطعی] است، نه آنها که به وضوح بدیع و جدید هستند؛ حالت اول، ریسک نظریه‌پردازی در مورد پدیده‌ای ناپایدار -وجه مشترک نامیمون بسیاری از نظریه‌های روز علوم اجتماعی - را کاهش می‌دهد. چهارم، ماهیت ذاتی نظریه علمی این است که پدیده‌ای پیچیده را به چیزی ساده‌تر تقلیل دهد، و نوسانات جزئی در اندازه اثر معمولاً ارتباط چندانی با ارکان نظریه ندارند. از آنجا که تعریف استحکام در قالب اندازه اثرهای ثابت، باعث از دست رفتن تمامی این مزایا می‌شود، این کتاب معیار سست‌تر – یعنی پایداری علامت علّی – را خصوصاً برای پژوهشهای کاربردی مطلوب‌تر می‌داند.
با این همه، کتاب حاضر معیار ثبات اندازه اثر را به طور کلی کنار نمی‌گذارد، چون برخی مواقع تفاوتهایی کوچک در اندازه اثر، کاربردهای نظری و کاربردی بزرگی دارند. نمونه‌ای از این حالت زمانیست که نتیجه موردنظر دربرگیرنده نوعی آسیب، مانند مرگ است. برای مثال، اگر اضافه شدن یک مهارکننده آنژیوژنز[۲۱۵] به شیمی‌درمانی می‌تواند امید به زندگی را به مدت شش ماه در بیماران مبتلا به سرطان پروستات افزایش دهد، و دارو ارزان بوده، و اثرات جانبی چندانی نداشته باشد، بسیاری از بیماران و پزشکان آنها -به دلیل ارزشی که برای زندگی کمی طولانی‌تر قائلند- خواستار اضافه شدن این دارو به برنامه درمان خواهند بود. در چنین قضاوتهایی، باید هرگونه تفاوت در ارزش نسبت داده‌شده به نوسانات ناچیز در اثر، تخمین هزینه‌های زمینه‌ای و منافع مداخلات، و آگاهی نسبت به عوارض جانبی مداخله لحاظ شود. پس مجدداً باید اذعان کرد که قضاوت در مورد روایی بیرونی یک رابطه علّی را نمی‌توان به اصطلاحی آماری تقلیل داد.

نمونه‌گیری تصادفی و روایی بیرونی

در این کتاب تأکید چندانی بر نمونه‌گیری تصادفی برای بدست آوردن روایی بیرونی صورت نگرفته است. دلیل اصلی این موضوع آن است که اینکار آنچنان پرهزینه است، که ندرتاً انجام آن برای آزمایشها توجیه‌پذیر است. اگرچه، برای مواقعی که توجیه‌پذیر باشد، انجام آن را قویاً توصیه می‌کنیم. زیرا همانطور که تخصیص تصادفی استنباطهای مرتبط با روایی درونی را تسهیل می‌کند، نمونه‌گیری تصادفی استنباطهای روایی بیرونی را تسهیل می‌نماید (البته با فرض نبود ریزش، همانطور که برای تخصیص تصادفی این فرض انجام می‌شود). برای مثال، اگر یک آزمایشگر قبل از اینکه افراد را به طور تصادفی به موقعیتها تخصیص دهد، آنها را به طور تصادفی

215 Angiogenesis inhibitor

نمونه‌گیری کرده باشد، این نمونه‌گیری تصادفی می‌تواند تضمین کند که -در محدوده خطای نمونه‌گیری- رابطه علّی متوسط مشاهده‌شده در نمونه، مشابه رابطه علّی متوسطی که در هر نمونه تصادفی دیگری (با همین اندازه) که از همین جمعیت گرفته شود، مشاهده خواهد شد، خواهد بود. همچنین تضمین‌کننده آن است که رابطه علّی متوسط مشاهده‌شده مشابه رابطه علّی متوسطی است که در میان تمامی دیگر افراد از جمعیت، که در نمونه تصادفی انتخاب نشده‌اند، نیز می‌توانست مشاهده شود. به این معنی که نمونه‌گیری تصادفی برهم‌کنشهای احتمالی میان رابطه علّی، و نوع افرادی که در نمونه انتخاب شده‌اند، و افرادی از همان جمعیت که در نمونه انتخاب نشده‌اند، را از میان می‌برد. اگرچه، این گونه آزمایشهای دارای نمونه‌گیری تصادفی نایاب هستند، در فصل ۱۱ مواردی از این نوع آزمایشها را به عنوان نمونه ارائه خواهیم نمود. علاوه بر این، فرض کنید که محقق بخواهد برهم‌کنش مداخله با یکی از مشخصات افراد (مثلاً جنسیت) را آزمون کند، باز نمونه‌گیری تصادفی تضمین خواهد کرد که برهم‌کنش در گروههای تعریف شده (۱) و (۲) یکسان خواهد بود. اگرچه تقسیم نمونه به گروه توان آماری آزمون را کاهش می‌دهد. بنابراین، با وجود اینکه در فصلهای اول و یازدهم محدودیتهای نمونه‌گیری تصادفی را در آزمایشها به تفصل مورد بحث قرار داده‌ایم، اما منافع آن برای تأمین روایی بیرونی به حدی است که در موارد معدودی که امکان انجام آن وجود دارد، حتماً می‌بایست از آن بهره برد.

انجام نمونه‌گیری تصادفی از [انواع] مختصات آزمایشی نیز به همین میزان سودمند است. مثلاً پوما و همکارانش(Puma et al., 1990) در آزمایشی با موضوع برنامه اشتغال و آموزش ضمن‌خدمت برنامه تغذیه کمکی، نمونه‌ای تصادفی از آژانسهای تغذیه کمکی انتخاب کردند. اما نمونه‌های تصادفی از مختصات در آزمایشها، حتی از نمونه‌های تصادفی افراد نیز کمیاب‌تر هستند. اگرچه جمعیت تعریف‌شده مختصات آزمایشها نسبتاً عمومی هستند (و در نتیجه در دسترس هستند) - مانند مراکز آموزش پیش‌دبستانی، مراکز سلامت‌روانی، یا بیمارستانها. اما نایابی نمونه‌گیریهای تصادفی از میان این جمعیتها احتمالاً به دلیل هزینه‌های لجیستیکی نمونه‌گیری تصادفی موفق از میان مختصات آزمایشی‌ست؛ هزینه‌هایی که باید به هزینه‌های بالای آزمایشهای چند-مکانی اضافه شوند.

در نهایت، این منافع (حاصل از نمونه‌گیری‌های تصادفی) برای مداخلهها و نتایج نیز صدق می‌کند. اما فهرست مداخله‌ها (Steiner & Gingrich, 2000)، و متغیرهای نتیجه‌ای (انجمن روانشناسان آمریکا، ۲۰۰۰) نیز کمیاب است، و هیچ تلاشی برای دفاع از اینگونه نمونه‌گیریهای تصادفی وجود ندارد. در مورد مداخلهها، این کمیابی به دلیل آن است که انگیزه انجام آزمایش در هر مطالعه‌ای، از سؤالات مرتبط با اثر یک مداخله خاص نشات می‌گیرد. و در مورد متغیرهای نتیجه‌ای، از آن روست که اغلب محققین احتمالاً بر این باورند که تنوع در مقیاسهای اندازه‌گیری نتایج، از طریق روشهای دقیق و حساب‌شده مانند موارد زیر به نحو بهتری حاصل می‌شود.

نمونه‌گیری هدفمند و روایی بیرونی

در مطالعه‌های تکی، استفاده از نمونه‌گیری هدفمند از موارد نامتجانس، در مقایسه با نمونه‌گیری تصادفی فراوانی بیشتری دارد. در نمونه‌گیری هدفمند، افراد، مختصات، مداخلهها، و نتایج به طور حساب شده‌ای انتخاب می‌شوند تا تنوع لازم را از نظر متغیرهایی که برای رابطه علّی اهمیت دارد، داشته باشند. برای مثال، اگر دلایلی دال بر اثر تعدیل‌کننده جنسیت بر رابطه‌ای وجود داشته باشد، باید نمونه مطالعه به صورتی انتخاب شود که دارای هر دو جنس زن و مرد باشد. انجام اینکار دو منفعت برای روایی بیرونی دارد. اول، اینکار امکان آزمون برهم‌کنش میان جنسیت و رابطه علّی را فراهم می‌سازد. اگر بتوان این برهم‌کنش را پیدا کرد، شاهدی بدیهی[۲۱۶] دال بر روایی بیرونی محدود در دست خواهیم داشت. اگرچه، برخی اوقات اندازه نمونه آنقدر کوچک است که آزمون برهم‌کنش میسر نیست، و در هر مطالعه تعدیلگرهای بالقوه متعددی وجود دارند که محقق به فکر آزمون آنها نمی‌افتد. در چنین موارد، بکارگیری نمونه‌گیری نامتجانس همچنان می‌تواند مفید باشد، چون می‌تواند نشان دهد که اثر یک مداخله، *علیرغم* عدم‌تجانس در نمونه همچنان برقرار است. یقیناً نمونه‌گیری تصادفی اثربخشی بیشتری برای انجام اینکار دارد، چون می‌تواند نمونه را از نظر تمامی متغیرهای تعدیلگر ممکن نامتجانس کند؛ اما کاربردی بودن نمونه‌گیری نامتجانس هدفمند، نقایصش را جبران می‌کند.

همین منافع از نمونه‌گیری هدفمند، از مختصات آزمایشی نامتجانس نیز بدست می‌آید. به نحوی که در تحقیقات چند-مکانی معمول است که محققین سطحی حداقلی از تنوع و ناهمسانی را تضمین کنند، به نحوی که دربرگیرنده مدارس عمومی و خصوصی یا بیمارستانهای خیریه و خصوصی باشد. نمونه‌گیری هدفمند از مقیاسهای اندازه‌گیری نتایج نامتجانس در اغلب حیطه‌های آزمایشهای میدانی چنان عمومیت دارد که ارزش آن برای کشف تعمیم‌پذیری اثرها محتوم فرض می‌شود؛ اگرچه نظریه‌های اندکی برای توضیح یا پیش‌بینی چنین تنوع‌پذیریهایی[۲۱۷] وجود دارد (Shadish & Sweeney, 1991). نمونه‌گیری هدفمند از مداخلههای نامتجانس در آزمایشهای تکی احتمالاً وجود ندارد، به همان دلایلی که نمونه‌گیری تصادفی از مداخلهها انجام نمی‌شود. اگرچه، در طول یک رشته برنامه‌های تحقیقاتی، و یا مجموعه‌ای از مطالعات که توسط محققین متفاوت انجام می‌شود، افراد، مختصات، مداخلهها، و نتایج به کرّات دارای سطح بالایی از تنوع و نامتجانسی خواهند بود. این یکی از دلایلی است که نظریه بنیادی تعمیم علّی در این کتاب تا حد زیادی بر روشهای چند-مطالعه‌ای تکیه دارد.

نکاتی دیگر در باب روابط، تعادلها، و اولویتها

در انتهای فصل دوم، رابطه میان روایی درونی و روایی نتایج آماری را مورد بحث قرار دادیم. در ادامه این بحث را بسط داده و دیگر روابط میان انواع روایی و اولویت و تعادل میان آنها را مورد بحث قرار خواهیم داد.

216 Prima facie
217 Variability

رابطه میان روایی سازه و روایی بیرونی

روایی بیرونی و روایی سازه به دو طریق با یکدیگر رابطه دارند. اول، هر دو نوعی تعمیم دادن هستند. در نتیجه، نظریه بنیادین تعمیم‌پذیری که در فصل اول معرفی کرده، و در فصول ۱۱ و ۱۳ به تفصیل مورد بحث و تفصیل قرار خواهیم داد، به خوبی می‌تواند ارتقاءدهنده هر دو نوع این روایی‌ها باشد. دوم، دانش معتبر و روا از سازه در یک مطالعه می‌تواند به روشن شدن سؤالات مرتبط با روایی بیرونی نیز کمک کند؛ علی‌الخصوص اگر نظریه‌ای محکم و خوش-ساخت وجود داشته باشد که توضیح دهد چطور سازه‌ها و موارد[۲۱۸] مختلف به یکدیگر ارتباط دارند. برای مثال، در پزشکی نظریه‌های محکمی در باب طبقه‌بندی درمانهای مشخص (مثل طبقه داروهایی که به آنها شیمی‌درمانی برای سرطان گفته می‌شود)، و اینکه چطور این درمانها بیماران را تحت تأثیر قرار خواهند داد (مثلاً چه تأثیری در آزمایش خون و ماندگاری بیماران داشته و چه عوارض جانبی‌ای دارد)، وجود دارد. در نتیجه هنگامی که یک دارو با توجه به معیارهای مربوطه در یک طبقه درمانی قرار می‌گیرد، می‌توانیم عمکلرد احتمالی آن را حتی قبل از امتحان کردن آن پیش‌بینی کنیم (برای مثال، می‌توانیم پیش‌بینی کنیم که احتمالاً مصرف دارو همراه با ریزش مو و اسهال بوده، و باعث افزایش طول عمر در بیماران با تومورهای کوچک، و نه در بیماران با تومورهای پیشرفته می‌شود). این دانش، طراحی آزمایشهای جدید را از طریق محدود کردن بیماران و نتایج آنها، ساده‌تر می‌کند، و باعث می‌شود تا برونیابیهای مرتبط با اثر مداخلهها با صحت بیشتری انجام شود. اما به مجرد اینکه به سمت اغلب موضوعات آزمایشهای میدانی حرکت می‌کنیم، با کمبود این نظریات خوش-ساخت و محکم مواجه می‌شویم. در این موارد، دانش حاصل از روایی سازه تنها شواهد ضعیفی درباره روایی بیرونی ارائه خواهد نمود. در فصلهای ۱۱ و ۱۳ مثالهایی درباره اینکه چطور این مسأله رخ می‌دهد، ارائه خواهیم کرد.

روایی سازه و روایی بیرونی بیش از آنکه شباهت داشته باشند، از جنبه‌های مختلف با یگدیگر تفاوت دارند. اول، آنها از نظر نوع استنباطهایی که ارایه می‌کنند، با یکدیگر تفاوت دارند. استنباط روایی سازه -بنا به تعریف- مفهومیست که همواره برای موارد مطالعه (اعم از متغیرها، مداخلهها، افراد و مختصات آزمایشی) بکار گرفته می‌شود. اما در تعمیم‌های روایی بیرونی، استنباطها با این مسأله سروکار دارند که آیا اندازه و جهت یک رابطه علّی برای افراد، مداخلهها، مختصات، و نتایج مختلف تغییر می‌کند یا نه. مثلاً یک مسأله چالش‌زا برای روایی سازه می‌تواند این باشد که مختصات آزمایش در یک مطالعه حوزه بهداشت را به اشتباه بیمارستانهای بخش خصوصی تعریف کرده‌ایم، در حالی که درست‌تر این بود که بگوییم بیمارستانهای خصوصی خیریه‌ای، تا آنها را از بیمارستانهای غیرانتفاعی خصوصی که موردنظر مطالعه نیستند، متمایز کرده باشیم. در هنگام طرح این چالش، هیچ اشاره‌ای به اندازه و جهت رابطه علّی نمی‌شود.

218 Instances

دوم، تعمیمهای روایی بیرونی را نمی‌توان از رابطه علّی تحت مطالعه جدا کرد، اما در مورد روایی سازه می‌توان این کار را انجام داد. این نکته در جمله‌بندی تهدیدات روایی بیرونی به وضوح مشهود است. این جملات همواره دربرگیرنده اثر برهم‌کنشی رابطه علّی با افراد، مداخلهها، مختصات و نتایج بالقوه است. مثلاً، هیچ تهدید روایی بیرونی که ناشی از اثر برهم‌کنش افراد و مختصات آزمایش، بدون ارتباط با رابطه علّی باشد، وجود ندارد. این به این معنی نیست که چنین برهم‌کنشهایی نمی‌توانند رخ دهند. مثلاً، همه می‌دانند که تعداد افراد با علائم مختلف بیماریهای روانی که می‌توان در یک بیمارستان روانی ایالتی یافت، با تعدادی که به طور کلی می‌توان در کلینیکهای روانپزشکی خصوصی پیدا کرد، کاملاً تفاوت دارد. حتی می‌توان در مورد روایی سازه برچسبها یا نامهایی که برای این سازه‌ها بکار بردیم نیز سؤالاتی مطرح کرد (برای مثال، آیا نام بیمارستانهای روانی ایالتی نام درستی است؟ یا شاید بهتر بود بگوییم اقامتگاههای ایالتی نگهداری بلند-مدت بیماران روانی. آیا این نام به نحو بهتری این اقامتگاهها را از مراکز ایالتی ارائه خدمات کوتاه مدت به بیماران روانی متمایز نمی‌ساخت؟). اما از آنجا که رابطه علّی در این برهم‌کنش خاص وجود ندارد، به روایی بیرونی ارتباط پیدا نمی‌کند.

البته در عمل، هنگام اشاره به تهدیدهای روایی از نامها و برچسبهای انتزاعی و خلاصه‌شده استفاده می‌کنیم. در دنیای واقعی علم، کسی نمی‌گوید که «من فکر می‌کنم این رابطه علّی برای افراد لیست الف برقرار است، اما در مورد افراد لیست ب صدق نمی‌کند». در عوض، ممکن است بگویند «فکر می‌کنم ژن-درمانی برای بیماران سرطانی با آسیب توموری کمتر، در مقایسه با افراد با تومورهای پیشرفته بهتر عمل می‌کند». اما استفاده از نام سازه‌ها به شکل دومی، باعث نمی‌شود روایی بیرونی با روایی سازه یکسان (و نه حتی به آن وابسته) باشد. مقایسه‌ای با روایی درونی در اینجا مفید به نظر می‌رسد. هیچ کسی در دنیای واقعی علم در مورد اینکه الف باعث ب می‌شود صحبت نمی‌کند. بلکه، همواره صحبت از روابط علّی توصیفی از منظر سازه‌هاست؛ مثلاً، گفته می‌شود ژن-درمانی ماندگاری بیمار را برای پنج سال افزایش می‌دهد. با این حال، در تعریف روایی درونی اینطور بیان شد که روایی درونی به این مسأله می‌پردازد که آیا الف بر ب اثر دارد؛ و هیچ اشاره‌ای به نام سازه نکردیم. علّت انجام این کار تأکید بر این حقیقت بود که الزامات و مسائل منطقی موجود در ارزیابی یک استنباط علّی توصیفی (آیا علّت پیش از اثر رخ داده، آیا دلایل جایگزین را می توان بی‌اثر ساخت، و غیره)، افضل و ارجح بر درستی نامگذاری آن سازه‌هاست. همین نکته برای روایی بیرونی نیز مصداق دارد. الزامات و مسائل منطقی موجود در ارزیابی اینکه آیا یک رابطه علّی برای افراد، مداخلهها، مختصات و نتایج مختلف برقرار است یا نه، بر مسایل مرتبط با نامگذاری سازه‌ها ارجحیت دارد.

سوم، روایی سازه و بیرونی از آنجهت با یکدیگر تفاوت دارند که ممکن است در مورد یکی قضاوتی درست، اما در مورد دیگری قضاوتی نادرست داشته باشیم. دو مجموعه از افراد را در نظر بگیرید که برای آنها عنوانهای سازه‌ای دقیقی و محکمی داریم، مثل عناوین مرد در مقابل زن، یا شهرهای آمریکایی در مقابل شهرهای کانادایی و یا

مقیاسهای خود-اظهاری در مقابل رتبه‌دهی مشاهده‌کننده. در این موارد، روایی سازه این برچسبها موضوع اصلی نسیت. باز در نظر بگیرید که آزمایشی با لحاظ کردن یکی از این مجموعه‌ها انجام دادیم، مثلاً تنها از مقیاسهای خود-اظهاری استفاده کردیم. این حقیقت که ما به درستی نام دسته دیگری که آن را در مطالعه وارد نساخته‌ایم را می‌دانیم، به ندرت می‌تواند برای پاسخ به این سؤال مرتبط با روایی بیرونی که، آیا اثر علّی مشاهده‌شده بر نتایج حاصل از خود-اظهاری، با نتایجی که ممکن بود با استفاده از نمره‌دهی مشاهده‌کننده بدست بیاید یکسان خواهد بود یا نه، کمکی خواهد کرد؟ (مواردی که در آن، نظریه‌های بسیار قوی برای پیش‌بینی روابط وجود دارد، از این موضوع مستثنی هستند). عکس این موضوع هم درست است. به این معنی که حتی زمانی که نامهای نادرستی برای این دو دسته از افراد بکار گرفته‌ایم، همچنان می‌توانیم پاسخهای مفیدی برای سؤالات روایی بیرونی، و اینکه آیا اثر موردنظر برای هر دو نوع نتایج -علیرغم نامگذاری اشتباه برای آنها- برقرار است، ارائه دهیم.

در نهایت، روایی بیرونی و سازه از نظر روشهایی که می‌توان برای ارتقاء هر یک از آنها بکار گرفت، متفاوت هستند. روایی سازه بیشتر بر توضیح روشن از سازه و ارزیابی درست از مشخصه‌های مطالعات استوار است، بگونه‌ای که بتوان در مورد تطابق و تناسب میان سازه و مشخصه‌های مطالعه قضاوت کرد. روایی بیرونی بیشتر بر بررسی تغییرات در اندازه و جهت رابطه علّی تکیه دارد. مطمئناً این بررسیها بدون همان ارزیابیها قابل‌انجام نیست، اما این در مورد روایی نتایج آماری و روایی درونی نیز صدق می‌کند؛ چون هر دو آنها در عمل به داشتن ارزیابی‌هایی که بتوان با آنها کار کرد، وابسته هستند.

رابطه میان روایی درونی و روایی سازه

روایی درونی و سازه در مفهوم متغیر کمکی[۲۱۹] مشترک هستند. رابطه روایی درونی و روایی سازه به بهترین نحو با چهار تهدیدی که برای روایی درونی ذکر شد (Cook & Campbell, 1979)، و در اینجا آنها را برای روایی سازه نیز مطرح می‌کنیم، تجلی پیدا می‌کند: تضعیف روحیه ناشی از رنجش و ناراحتی، همسان‌سازی جبرانی، رقابت جبرانی و انتشار مداخله. این مسأله که آیا این تهدیدها را باید ذیل روایی درونی و روایی سازه به حساب بیاوریم یا نه، منوط به آن است که بدانیم این تهدیدها دقیقاً چه نوع متغیرهای کمکی ایجاد می‌کنند. متغیرهای کمکی روایی درونی، نیروهایی هستندکه می‌توانستند در غیاب مداخله نیز رخ دهند، و نتایجی مشابه و یا کاملاً مطابق نتایج مداخله تولید کنند. در مقابل، اگر یک مداخله خاص اجرا نشده بود، این چهار تهدید مجال بروز پیدا نمی‌کردند. این تهدیدها به این دلیل رخ داده‌اند که مداخله اعمال شده، و بنابراین آنها نیز بخشی از شرایط مداخله به حساب می‌آیند (یا به بیان دقیقتر، بخشی از نقیضهای[۲۲۰] مداخله هستند). این تهدیدها، روایی سازه را

219 Confound
220 Contrasts

به این دلیل به خطر می‌اندازند که بخشی از ساختار مفهومی موردنظر مداخله نیستند، و بنابراین، اغلب از توصیف سازه مداخله حذف می‌شوند.

تاخت‌زنی‌ها[۲۲۱] و اولویتها

در دو فصل گذشته فهرستی بلند از تهدیدات متوجه روایی استنباطهای علّی تعمیم‌یافته را ارائه کردیم. این ممکن است باعث شود خوانندگان با خود فکر کنند که آیا آزمایشی وجود دارد که بتواند از تمامی این تهدیدها اجتناب کرده و مبرا باشد؟ پاسخ منفی است. به طور منطقی نمی‌توانیم انتظار داشته باشیم که یک مطالعه بتواند همزمان تمامی این تهدیدات را رفع نماید، چون همواره تاخت‌زنیهای لجیستیکی و عملی میان این تهدیدات انجام می‌گیرد. تاخت‌زدنهایی که در این فصل به آن بیشتر خواهیم پرداخت. [لیست] تهدیدات روایی، ابزارهایی شهودی برای افزایش هوشیاری نسبت به اولویتها و تاخت‌زنیها هستند، و نه منبی برای بدبینی و یأس. برخی از تهدیدها از نظر اولویت و عواقب آنها بر کیفیت استنباطها با اهمیت‌تر از دیگر تهدیدها هستند، و محققین به تجربه درمی‌یابند آنهایی را که اهمیت بیشتر داشته و برای یک زمینه خاص اولویت بیشتری دارند، را شناسایی کنند. معقول‌تر آن است که از یک پروژه پژوهشی انتظار داشته باشیم بتواند در طول زمان بر اغلب یا تمامی این تهدیدها فائق بیاید. رشد دانش بیشتر فرایندی تجمعی و انباشتی است تا اپیزودیک [که در یک تحقیق حاصل شود]، خواه در آزمایشها و خواه دیگر انواع پژوهشها. اگرچه منظور از بیان این مطالب این نیست که یک آزمایش منفرد بی‌فایده بوده، و یا نتایج تمامی آنها به یک اندازه مشحون از عدم‌اطمینان است. برای داشتن یک آزمایش خوب لازم نیست تمامی تهدیدات را برطرف کنیم، بلکه یک آزمایش خوب باید بتواند زیرمجموعه‌ای از تهدیدات که زمینه خاصی از مطالعه را بطور جدی به مخاطره می‌اندازند را مورد توجه قرار دهد. گذشته از آن، مواجهه با تهدیدات و برطرف کردن آنها تنها نشانه یک آزمایش خوب نیست؛ به طور مثال، بهترین آزمایشها آنهایی هستند که می‌توانند با آزمودن ایده‌هایی بدیع و تازه زمینه‌های مختلف تحقیقاتی را تحت تاثیر قرار دهند (Eysenck & Eysenck, 1983; Harre, 1981).

در دنیای محدودیت منابع، محققین در یک مطالعه منفرد، همواره میان انواع روایی تاخت‌زنی و بده‌بستان می‌کنند. به عنوان نمونه، اگر محققی برای بهبود روایی نتایج آماری اندازه نمونه را افزایش دهد، در واقع منابعی را که می‌توانستند برای جلوگیری از ریزشهای همبسته با مداخله مورد استفاده قرار گرفته، و در نتیجه روایی درونی را بهبود دهند، را کاهش داده است. به همین صورت، تخصیص تصادفی به میزان زیادی به افزایش روایی درونی کمک می‌کند؛ اما سازمانهایی که مایل به تحمل هزینه‌های انجام آن هستند، به اندازه سازمانهایی که به انجام اندازه‌گیریهای منفعل بسنده می‌کنند، نماینده [جامعه] نیستند، بنابراین روایی بیرونی احتمالاً مورد

221 Tradeoff

مسامحه قرار می‌گیرد. همچنین افزایش روایی سازه اثرها از طریق عملیاتی کردن سازه‌ها به شیوه‌های متعدد، باعث [طولانی شدن پرسشنامه] و دشواری پاسخگویی به پرسشنامه و در نتیجه، ریزش پاسخ‌دهندگان می‌شود. یا اگر بودجه محاسباتی ثابت باشد، افزایش تعداد مقیاسها می‌تواند پایایی هر یک از این مقیاسها را - که به ناچار باید کوتاه شوند - کاهش دهد.

چنین روابط متناقضی نشان‌دهنده اهمیت ذکر ترتیب اهمیت انواع مختلف روایی در هنگام طراحی یک آزمایش است. باید از تسامحها و تاخت‌زدنهای غیرضروری میان انواع مختلف روایی اجتناب کرد، و خسارت ناشی از بده‌بستانهای ضروری نیز باید تخمین‌زده شده و به حداقل رسانده شود. محققین مختلف تخمینهای متفاوتی از موازنه‌های مطلوب دارند. کرونباخ (Cronbach, 1982) بر این باورست که مطالعاتی که به روز بوده، و نماینده جامعه باشند، اما از قوام و استحکام کمتری برخوردار هستند، می‌توانند به استنباطهای علّی ارزشمندی منتهی شوند که از روایی بیرونی خوبی برخوردار باشند، حتی اگر این مطالعات به طور غیرآزمایشی انجام شوند. از سوی دیگر، کمپبل و براخ (Campbell & Boruch, 1975) بر این باورند که روابط علّی بیرون از آزمایشها مسأله‌دار است؛ چون بسیاری از تهدیدات متوجه روایی درونی ارزیابی نشده باقی می‌ماند، و یا اینکه به جای آنکه در جریان طراحی مستقیم آزمایش یا اندازه‌گیری بی‌اثر شوند، باید بصورت دستوری از چرخه اثرات خارج شوند. این تاخت‌زنی میان روایی درونی و بیرونی، نمونه‌ای از مهترین و عمده‌ترین تاخت‌زنیها میان انواع روایی ست.

روایی درونی: شرطی ضروری

با توجه به اینکه روایی درونی و بیرونی در اغلب مطالعات در تقابل با یکدیگر قرار می‌گیرند، کمپبل و استنلی (Campbell & Stanley, 1963) می‌گویند: «روایی درونی شرطی ضروریست[۲۲۲]»(ص ۵). این جمله به روایی درونی در نسلی از آزمایشهای میدانی اولویت و ارجحیت می‌دهد. اما کرونباخ با این اولویت مسأله دارد. او بر این باورست که روایی درونی «سطحی، مربوط به زمان گذشته، و محلی است »(۱۹۸۲، ص ۱۳۷)، در حالی که روایی بیرونی نگاه به جلو داشته بوده و سؤالات کلی و عمومی می‌پرسد، در نتیجه از اهمیت بیشتری برخوردار است. از آنجا که این نگاه نسبت به یک گونه‌شناسی اصیل برای روایی تنها نظر کرونباخ نیست، در اینجا به بحث در مورد اولویت روایی درونی بر دیگر روایییها علی‌الخصوص روایی بیرونی خواهیم پرداخت.

این گفته کرونباخ و استنلی (۱۹۶۳) که روایی درونی شرط ضروری آزمایش است، یکی از جملاتی است که در روش تحقیق بیشترین ارجاع به آن داده شده است. این جمله در کتابی با موضوع آزمایش و شبه‌آزمایش مطرح

222 Sine Qua Non (چیزی که بدون آن، انجام کاری امکانپذیر نیست)

شد، و در متن به وضوح اشاره شده است که این جمله منحصراً آزمایشها را منظور نظر داشته، و دیگر شکلهای مطالعات را در بر نمی‌گیرد:

روایی درونی حداقلِ لازم است، که بدون آن، هر آزمایشی غیرقابل تفسیر خواهد بود: آیا مداخلههای آزمایشی در حقیقت در این زمینه آزمایشی خاص تفاوتی ایجاد کرده‌اند؟ روایی بیرونی سؤالاتی درمورد امکان تعمیم مطرح می‌کند: این اثر را به چه جمعیتها، مختصات، متغیرهای مداخله و متغیرهای اندازه‌گیری می‌توان تعمیم داد؟ هر دو نوع این معیارها یقیناً مهم هستند. اگرچه مکرراً در تقابل با یکدیگر قرار داشته، و عناصری که باعث تقویت یکی می‌شود، دیگری را تضعیف می‌کند. با وجود آنکه روایی درونی شرطی ضروری بوده، و پاسخ به سوال روایی بیرونی (مثل سؤال در مورد استنباط استقرایی) هیچگاه به طور کامل میسر نیست، انتخاب طرحهایی که از نظر هر دو روایی (بیرونی و درونی) قوی باشند ایده‌ال است. این مسأله خصوصاً در مورد تحقیقات حوزه آموزش که در آن تعمیم به شرایط یا مختصات عملی مرتبط با نوع خاصی از شخصیت بسیار مطلوب[۲۲۳] است، اهمیت بیشتری می‌یابد (Campbell & Stanley, 1963, p.5).

در نتیجه، استنلی و کمپبل بر این باورند که روایی درونی برای طرحهای آزمایشی یا شبه‌آزمایشی که فرضیات علّی را مورد بررسی قرار می‌دهند، ضرورت دارد، و نه برای همه انواع مطالعات به عنوان یک قائده عمومی. بعلاوه، آخرین جمله این نقل قول تقریباً همواره نادیده گرفته می‌شود. این جمله اشاره می‌کند که روایی بیرونی یک آرزو و مطلوب (هدف، قصد، ملزوم) در مطالعات حوزه آموزش است. این جمله نیز چیزی به قوت جمله گفت‌شده در مورد روایی درونیست. همانطور که کوک و کمپبل (Cook & Campbell, 1979) در ادامه بیان می‌کنند، عبارت شرط ضروری تا اندازه‌ای حشو و زائد است:

همچنین نوعی توجیه دایره‌وار[۲۲۴] برای روایی درونی وجود دارد، که در هر کتاب مرتبط با آزمایش می‌توان آنرا یافت. هدف منحصر به فرد آزمایشها، تأمین آزمونهایی قویتر برای فرضیات علّیست. بررسیهایی که با دیگر انواع تحقیق که برای اهداف دیگر طراحی شده‌اند امکان‌پذیر نیست. برای مثال، پیمایشها برای تبیین نگرش جمعیت و رفتارهای گزارش شده طراحی شده است، در حالی که روشهای مبتنی بر مشاهده شرکت‌کنندگان برای توضیح و تولید فرضیات جدید درمورد رفتارهای در حال انجام در محل اصلی خود، ساخته شده است. با توجه به اینکه هدف اصلی آزمایشها علّت-محور است، در آزمایش باید اهیمتی مضاعف برای روایی درونی قائل شویم؛ چون روایی درونی به این می‌پردازد که تا چه اندازه

223 Desideratum
224 Circular justification

می‌توانیم مطمئن باشیم که رابطه مشاهده شده میان متغیرها علّی است و یا اینکه نبود یک رابطه نشان‌دهنده نبود علّت است (ص ۸۴).

علیرغم تمامی این تکذیبها، بسیاری از خوانندگان، موضع ما نسبت به روایی درونی را به درستی متوجه نمی‌شوند. برای جلوگیری از این سوءبرداشتها، اجازه بدهید به طور شفاف بگوییم: *روایی درونی شرط ضروری تمامی انواع تحقیق نیست (شرطی که بدون آن، انجام تحقیق غیرممکن باشد)، بلکه بواسطه تأکیدی که بر تفکر انتقادی درباره ادعاهای علّی توصیفی می‌نماید، جایگاهی ویژه (اما نه مقدس و غیرقابل نقض) در تحقیقات علّت‌یابی و علی‌الخصوص در تحقیقات آزمایشی دارد.* در ادامه برخی از مسائل که می‌باید پیش از دانستن میزان اولویت روایی درونی بدانیم را مورد بررسی قرار می‌دهیم.

آیا علیّت توصیفی یک اولویت است؟

روایی درونی تنها زمانی می‌تواند از اولویت و ارجحیت بالا برخوردار باشد که محققی از میان سؤالات رقیبی که می‌تواند درخصوص یک موضوع خاص پرسیده شود، آگاهانه علاقمند به دانستن در مورد سؤال توصیفی علّی خاصی باشد. سؤالات رقیب می‌تواند این باشد که مسأله چطور صورتبندی شده است؟ مداخله چه نیازهایی را برطرف می‌کند؟ مداخله موردنظر چقدر خوب اجرا (اعمال) شده است؟ چطور می‌توان چیزی را به بهترین نحو اندازه‌گیری کرد؟ چطور فرایندهای واسطه‌گر را باید شناخت؟ چطور باید یافته‌ها را معنا کرد؟ و منافع و هزینه مالی را چطور باید محاسبه شود؟ به ندرت در آزمایشها اطلاعات مفیدی در ارتباط با این سؤالات ارائه می‌شود، سؤالاتی که دیگر روشها برای پاسخگویی به آنها ارجح است. حتی زمانی که علّیت توصیفی یک اولویت باشد، لازم است تا، در حیطه همان منابع محدود تحقیق، دیگر سؤالات فوق‌الذکر نیز پاسخ گفته شوند. در نتیجه ممکن است روشی مانند پیمایش بهتر باشد چون گستره دامنه[۲۲۵] وسیعتری داشته، و در نتیجه امکان پاسخ دادن به طیف وسیعتری از سؤالات را فراهم می‌آورد، حتی اگر پاسخ سوال علّی به آن خوبی که از طریق یک طرح آزمایشی داده می‌شود، نباشد (Cronbach, 1982). تصمیم برای اولویت دادن به سوالات علّی توصیفی یا دیگر سوالات فراتر از این کتاب است (Shadish, Cook, & Leviton, 1991). فرض ما بر این است که محقق قبل از شروع به کار در چهارچوب آزمایشی که مورد بحث در این کتاب است، به پاسخ این سوال رسیده است.

آیا روشهای غیرآزمایشی می‌توانند پاسخهایی راضی‌کننده ارائه کنند؟

225 Bandwidth

حتی اگر یک استنباط علّی توصیفی به عنوان یک اولویت اصلی به خوبی توجیه‌پذیر باشد، روشهای آزمایشی همچنان تنها گزینه نیستند. سؤالات توصیفی علّی را می‌توان غیرآزمایشی نیز بررسی کرد. این موضوع در مطالعات تحلیل مسیر در جامعه‌شناسی(Wright, 1921, 1934)، مطالعات مورد-کنترل در اپیدمیولوژی (Ahlbom & Norell, 1990) و یا روشهای کیفی مانند مطالعات موردی (Campbell, 1975) اتفاق می‌افتد. تصمیم برای بررسی یک سوال توصیفی علّی با چنین روشهایی به عوامل متعددی بستگی دارد. این تا حدی انعکاس‌دهنده سنتهای هر رشته است، سنتهایی که با دلایلی خوب یا ضعیف شکل گرفته‌اند. برخی پدیده‌ها غیرقابل‌دستکاری هستند- چیزی که لازمه مطالعات آزمایشی ست. در مورد برخی دیگر، انجام دستکاری غیراخلاقیست، و یا انجام دستکاری ممکن است پدیده موردنظر را به نحو غیرمطلوبی تغییر دهد. بعضی مواقع علّت موردنظر هنوز به طور کامل روشن نشده، و بنابراین بیشتر علاقمندیم که طیفی از علّتهای ممکن را کشف کنیم تا اینکه روی یک یا دو علّت تمرکز کنیم. برخی اوقات هنوز زمان اختصاص وقت و منابع به آزمایشها نرسیده، شاید چون فاز پایلوت انجام‌شده برای طراحی یک مداخله، از نظر صحت نظری و اجرایی بودن عملی ناکافی بوده، و یا چون جنبه‌های مهم فرایند آزمایشی مانند محاسبه نتایج به طور ناقص طراحی شده، و یا شاید نتایج را به سرعت احتیاج داشته، و نمی‌توانیم زمان لازم را برای انجام آزمایش را اختصاص دهیم. کارهای آزمایشی تکامل‌نیافته و پیش از موعد یکی از گناهان معمول پژوهشی است.

اگرچه، ماهیت روشهای غیرآزمایشی مانع از آن می‌شود که بتوانند بالاترین اولویت را به روایی درونی اختصاص بدهند. علّت آن است که روشهای آزمایشی بیش از دیگر روشها با الزامات استدلالهای علّی تناسب دارند؛ یعنی در تضمین این سه شرط که علّت پیش از اثر اتفاق می‌افتد، منبع معتبری از استنباط خلاف‌واقع وجود دارد، و تعداد گزینه‌های جایگزین موجّه کاهش داده شده است. با این حال، داده‌های مورد استفاده در روشهای علّی غیرآزمایشی، در مقایسه با روشهای آزمایشی اغلب از نمونه‌هایی بدست می‌آیند که نماینده بهتری برای سازه‌ها هستند. همچنین، طرح‌های غیرآزمایشی اغلب طرحهای نمونه‌گیری وسیعتری دارند که روایی بیرونی را به نحو بهتری تسهیل می‌کند. بنابراین در مقایسه با روشهای آزمایشی، روشهای غیرآزمایشی غالباً قابلیت کمتری برای تأمین روایی درونی دارند، اما توانایی برابر و یا حتی برتر برای تأمین روایی بیرونی و روایی سازه دارند. البته این قابلیتها عمومی و فراگیر نبوده، و مواقعی پیش می‌آید که روشهای غیرآزمایشی، استنباطهای توصیفی علّی‌ای با کیفیتی برابر با استنباطهای بدست‌آمده از آزمایشها بدست می‌دهند (مانند استنباطهای مطالعات اپیدمیولوژی). همانطور که در ابتدای فصل دوم اشاره کردیم، روایی خصلت ادعاهای دانشی[۲۲۶] است نه روشها. داشتن روایی درونی بیشتر وابسته به برآورده کردن الزامات استدلال علّیست، تا استفاده از یک روش خاص. هیچ روشی، حتی آزمایش، نمی‌تواند تضمین‌کننده حصول یک استنباط علّی با روایی درونی باشد، اگرچه که آزمایشها نسبت به بقیه در این خصوص برتری دارند.

جنبه‌های قوی و ضعیف شرط ضروری

فرض کنید که محقق در مورد تمامی موارد بالا بیاندیشد، و تصمیم بگیرد که از روش آزمایشی برای مطالعه یک استنباط علّی توصیفی استفاده کند. در این حالت، روایی درونی از دو وجه شرط ضروری خواهد بود. وجه ضعیف آن همان جنبه ایست که به زعم کمپبل و استنلی زائد یا حشو است: «روایی درونی حداقل پایه‌ایست که بدون آن، هیچ آزمایشی قابل‌تفسیر نیست (ص ۵)». به این معنا که انجام آزمایش و توجه نداشتن به روایی درونی نوعی پارادکس یا ضد و نقیض‌گویی است. انجام آزمایش تنها زمانی معنادار است که محقق علاقمند به سؤالات علّی توصیفی باشد، و چنین علاقه‌ای را بدون توجه همزمان به روایی پاسخ علّی، به سختی می‌توان توجیه و درک کرد.

جنبه قویتری که در آن روایی درونی ارجحیت و اولویت می‌یابد، زمانی رخ می‌دهد که محقق می‌تواند در یک آزمایش انتخاب کند چه میزان اولویت و اهمیت باید به هر یک از انواع روایی اختصاص یابد. متأسفانه هر تلاشی برای پاسخ به این سؤال بسیار پیچیده است، چون (۱) هیچ مقیاس پذیرفته‌شده‌ای که مقدار مناسب هر یک از این روایی‌ها را تعیین نماید نداریم؛ و (۲) بسیار دشوار است که تعیین کنیم چه مقدار از هر کدام از این روایی‌ها در حال حاضر وجود دارد. یک راه یا گزینه، استفاده از شاخصهای روش‌شناختی است. مثلاً اینکه ادعا کنیم مطالعات تصادفی با نرخ پایین ریزش استنباطهایی بدست می‌دهند که به احتمال زیاد از نظر روایی درونی در سطح بالاتری قرار دارند. اما شاخصی نمی‌تواند مقدار روایی درونی دیگر مطالعات علّت‌یابی را اندازه‌گیری کند. گزینه دیگر، استفاده از مقیاسها بر مبنای تعداد تهدیدهای شناسایی شده‌ای است که همچنان لازم است بی‌اثر شوند. اما موانع مفهومی بر سر راه این مقیاسها جدیست. حتی اگر بتوانیم مقیاسهایی را برای هر یک از انواع روایی بسازیم، هیچ راهی برای معرفی خط‌کشی مشترک که با آن بتوان اولویتها را مقایسه کرد، وجود ندارد. یک گزینه قابل‌اعتنا، بکارگیری مقدار منابع لازم برای تأمین هر یک از روایی‌ها، به عنوان شاخصی از اولویت و اهمیت خواهد بود. می‌توان با صرفه‌جویی منابع اختصاص‌یافته به روایی درونی، مابقی منابع را برای تأمین برخی دیگر از انواع روایی بازتوزیع کنیم. مثلاً محقق می‌تواند بخش از منابعی را که به روایی درونی اختصاص پیدا می‌کرد را به اندازه‌گیری سوگیری انتخاب، یا کاهش ریزش اختصاص دهد، و یا آنها را برای (۱) مطالعه تعداد بیشتر افراد (برای تسهیل روایی نتایج آماری)؛ (۲) اجرای شبه‌آزمایشهای متعدد روی مداخلههای موجود در مکانهای بیشتری که نماینده [جمعیت] باشند (برای ارتقاء روایی بیرونی)؛ (۳) افزایش کیفیت محاسبه نتایج (به منظور تسهیل روایی سازه)، صرف نماییم. این شکل از تخصیص منابع به طور اثربخشی اولویت روایی درونی را کاهش می‌دهد.

226 Knowledge claim

تصمیم‌گیری در مورد تخصیص منابع به متغیرهای بسیاری وابسته است. یکی از این عوامل، تمایز میان تحقیقات پایه‌ای و کاربردی است. تحقیقات بنیادی اهیمت زیادی به روایی سازه می‌دهند. چون سازه‌ها نقش بسیار مهمی در شکل‌دهی به نظریه‌ها و آزمون آنها دارند. محققین تحقیقات کاربردی اما بیشتر به روایی بیرونی توجه دارند؛ چون دانستن اینکه نتایج علَی بدست آمده، به زمینه کاربرد موردنظر نیز قابل تعمیم است یا نه، از اهمیت ویژه‌ای در این تحقیقات برخوردار است. برای مثال، آزمایشهای فستینگر (Festinger, 1953) در روانشناسی اجتماعی پایه به این خاطر مشهور هستند که محقق دقت فراوانی کرد تا مطمئن شود که متغیری که دستکاری‌شده دقیقاً همان «ناهماهنگی شناختی» بوده است. به همین ترتیب، در رابطه با افراد موردآزمایش، روانشناسان رشد پیرو پیاژه، اغلب منابع قابل‌توجهی را به این اختصاص می‌دهند که ارزیابی کنند کودک موردمطالعه در مرحله پیش-عملیاتی[۲۲۷]رشد است یا عملیاتی عینی[۲۲۸]. در مقابل، روایی سازه مختصات آزمایش از اهمیت کمتری در تحقیقات بنیادی برخوردار است، چون کمتر نظریه‌ای مختصات هدف را مشخصاً تعریف می‌کند. و در نهایت، غالباً کمترین اهمیت و توجه در تحقیقات بنیادین، به روایی بیرونی داده می‌شود. بسیاری از تحقیقات پایه‌ای روانشناسی با استفاده از دانشجویان سال‌های اول و دوم در دانشکده‌های روانشناسی انجام می‌شوند. علّت اینکار امکان داشتن تعداد زیاد پاسخ‌دهندگان یکدست است که باعث افزایش توان آماری می‌شود. این تاخت‌زدن با این امید صورت می‌گیرد که نتایج بدست‌آمده با این نوع دانشجویان کلی خواهد بود، چون این دانشجویان با استفاده از فرایندهای عمومی روانشناختی عمل خواهند کرد (پیشفرضی که نیازمند آزمونهای مکرر تجربیست[۲۲۹]). اگرچه، با فرض اینکه این مثالها در یک زمینه آزمایشی انجام شده‌اند، همچنان بسیار غیرمحتمل است که محقق اجازه داده باشد که منابع اختصاص‌یافته به روایی درونی، از سطحی حداقلی کمتر شده باشد.

227 Preoperational
228 Concrete operational
229 Empirical

بالعکس، بسیاری از تحقیقات کاربردی اولویتهای متفاوتی دارند. آزمایشهای کاربردی اغلب به دنبال بررسی این موضوع هستند که آیا مداخله پیشنهادی می‌تواند مرهمی برای مشکل خاص موردنظر آنها باشد؟ برای مثال، بحثهای مطرح در مورد اینکه کدامیک از انواع تعدیل هزینه‌های زندگی[۲۳۰] مبتنی بر شاخص قیمت مصرف‌کننده (CPI) به بهترین نحو نشان‌دهنده افزایش هزینه‌های زندگی است، را در نظر بگیرید. و یا اینکه آیا اساساً باید شاخص قیمت مصرف‌کننده (CPI) را به عنوان مقیاس هزینه‌های زندگی در نظر گرفت؟ به همین طریق، محققین رواندرمانی (Jacobson, Follette, & Revenstorf, 1984) این موضوع را مورد بررسی قرار داده‌اند که آیا مقیاسهای [سنجش] نتایج درمانهای سنتی به درستی انعکاس‌دهنده مفهوم «ارتقاء کلینیکی قابل‌توجه» در میان مراجعین کلینیکی خواهند بود؟ همچنین تحقیقات کاربردی اهمیت فراوانی برای تعمیم منطقی و معقول به اهداف روایی بیرونی که موضوع توجه جامعه [محل] کاربرد هستند، قائل است. برای مثال، وایز، وایس و دوننبرگ (Wiesz, Weiss, & Donenberg, 1992) بر این باورند که اغلب آزمایشهای رواندرمانی با استفاده از افراد، مداخلهها، مشاهدات، و مختصاتی انجام می‌شوند که بسیار از موارد مورد استفاده در کلینیکها فاصله دارند. این باعث می‌شود تا روایی بیرونی استنباطهای مرتبط با اینکه رواندرمانی تا چه اندازه در این زمینه‌ها خوب عمل خواهد کرد، کاهش یابد.

این مثالها نشان می‌دهند که تصمیمات در مورد اولویت و اهمیت نسبی انواع مختلف روایی در یک آزمایش را نمی‌توان در خلاء اتخاذ کرد. بلکه محققین باید موقعیت دانش در آن زمینه مطالعاتی را در نظر بگیرند. برای مثال، در «مدل فازبندی[۲۳۱]» تحقیقات سرطان در موسسه ملی بهداشت (Greenwald & Cullen, 1984)، استنباطهای علّی درباره اثرات مداخلهها همواره یک مسأله است، اما در فازهای مختلف مطالعه انواع مختلفی از روایی اهمیت و اولیت پیدا می‌کنند. در مراحل اولیه، تحقیق مرتبط با مداخلههای اثربخش احتمالی، طراحی‌های آزمایشی ضعیفتری دارند، که این امر مجال بروز بسیاری از نتایج به اشتباه مثبت[۲۳۲] (یا همان خطای نوع دو) را فراهم می‌کند (از ترس آنکه مبادا یک مداخله اثربخش احتمالی را نادیده بگیرند). همزمان با رشد بیشتر دانش، روایی درونی اولویت بیشتری می‌یابد، و لازم است بتوانند مداخلههایی را شناسایی کنند که دست کم تحت شرایط ایده‌ال اثربخشی واقعی دارند (مطالعات کارآمدی[۲۳۳]). در فاز نهایی تحقیق روایی بیرونی و اینکه مداخله تا چه اندازه تحت شرایط کاربرد در محیط واقعی درست عمل خواهد کرد، اولویت پیدا می‌کند (مطالعات اثربخشی).

230 Cost-of-living adjustment
231 Phase model
232 False positives
233 Efficacy studies

برنامه‌های تحقیقاتی نسبتاً کمی تا به این اندازه نظام‌مند هستند. اگرچه ممکن است این چهار نوع روایی به نظر راهنمای ضعیفی برای برنامه‌های آزمایشی باشد، توصیه می‌شود همزمان با مشخص شدن ضعفهای نسبی هر یک از رواییها در دانش علّی تعمیم‌یافته، محققین به طور رفت و برگشتی میان انواع مختلف روایی چرخش کنند. برای مثال، بسیاری از محققین با دیدین یک رابطه جالب میان دو متغیر یک برنامه پژوهشی را آغار می‌کنند (McGuire, 1997). آنها احتمالاً مطالعات بیشتری برای تأیید اندازه و قابل اتکاء بودن رابطه (روایی نتایج آماری) انجام می‌دهند، و سپس بررسی می‌کنند که آیا رابطه علّی است یا نه (روایی دورنی)، و در مرحله بعد، با دقت بیشتری آن را تعریف کنند (روایی سازه)، و حیطه‌ها و محدوده‌های آن را تعیین می‌کنند (روایی بیرونی). برخی اوقات، پدیده‌ای که کنجکاوی محقق را به خود جلب کرده، فی‌النفسه روایی بیرونی قابل‌توجهی دارد. به عنوان نمونه، کوواریانس میان سیگار کشیدن و سرطان ریه در افراد مختلف، در شرایط مختلف، و در طول زمانهای متفاوت، باعث تعریف یک برنامه تحقیقات آزمایشی برای (۱) تعیین اینکه رابطه مشاهده شده علّیست یا نه، (۲) تعیین اندازه اثر و وابستگی، و (۳) توضیح رابطه، طراحی شده. در مواردی دیگر، روایی سازه متغیرها همواره مورد توجه مطالعه بوده، اما به یکباره سؤال درباره یک رابطه علّی میان متغیرها تمامی توجه‌ها را به خود جلب می‌کند. بطور مثال، در سال ۱۹۹۰ بحث در مورد رابطه میان نژاد و هوش بالا گرفت، در حالی که روایی سازه این دو سازه به طور گسترده‌ای تا پیش از آن موردمطالعه قرار گرفته بود (Devlin, 1997; Herrnstein & Murray, 1994). برنامه‌های تحقیقات آزمایشی ممکن است در زمانهای متفاوت، با انواع مختلف نقاط‌قوت در دانش موجود که به استنباطها قوت ببخشند، و با نیاز به ترمیم انواع مختلف نقاط ضعف در دانش موجود، آغاز شوند. در طول یک برنامه تحقیقاتی، تمامی انواع روایی‌ها اهمیت دارند؛ و می‌باید تا پایان پروژه تحقیقاتی، هر کدام از آنها به نوبه خود در مرکز توجه قرار بگیرند.

فصل چهارم

طرحهای شبه‌آزمایشی فاقد گروه کنترل و یا مشاهدات پیش‌آزمون

Quasi ، از ریشه کلمه لاتین quasi: quam ، به معنای شبیه یا مانند بودن، و یا به چیزی شباهت داشتن است.

این فصل را با یک مثال که نشان‌دهنده نوع طراحیهای مورد بحث در این فصل است، آغاز می‌کنیم. در سال ۱۹۶۶ به منظور مقابله با عقب‌ماندگی ذهنی ناشی از فنیل کتونوریا (PKU)، برنامه‌ای برای پایش و درمان نوزادان بدنیا آمده با این بیماری در استان آنتوریو کانادا کلید خورد. ارزیابی انجام شده پس از اجرای برنامه نشان داد که ۴۴ نوزاد بدون عقب‌ماندگی ذهنی و تنها ۳ نوزاد علائم عقب‌ماندگی ناشی از PKU را نشان می‌دادند (از این سه نفر، دو نفر از برنامه پایش جامانده بودند). آمارهای بدست آمده از سالهای قبل، نشان‌دهنده نرخ بالاتری از عقب‌ماندگی ناشی از PKU بود. اگرچه روش انجام این مطالعه بسیار ابتدایی بود، چون این مطالعات گروه کنترلی[۲۳۴] که درمان را دریافت نکرده باشند نداشت، با این حال محققین چنین نتیجه‌گیری کردند که این برنامه توانسته به طور موفقیت‌آمیزی از عقب ماندگی ذهنی ناشی از PKU پیشگیری کند. در پی این نتایج، برنامه‌های مشابه به طور گسترده‌ای در کانادا و آمریکا به کار گرفته شد، و همچنان اثربخشی قابل‌توجهی از خود

[۲۳۴] عبارت گروه کنترل معمولاً به گروهی اطلاق می شود که مداخله یا درمان را دریافت نمی کند؛ عبارت کلی تر گروه مقایسه (comparison group) در بر گیرنده گروه کنترل و گروههای دریافت کننده مداخله ها و درمانهای جایگزین است.

نشان داد. چطور این مطالعه علّی رغم اینکه نه گروه کنترل داشت، و نه تخصیص تصادفی، توانست چنین نتایج روشن، درست و مفیدی بدست آورد؟ فصل پیش رو به این موضوع می‌پردازد.

ما در این فصل به توضیح طرحهای شبه‌آزمایشی که یا گروه کنترل ندارند، و یا فاقد مشاهدات پیش‌آزمون در مورد نتایج هستند، خواهیم پرداخت. اگرچه مطالعات محدودی از این دست توانسته‌اند نتایج علّی روشنی مانند آنچه در مثال بالا به آن اشاره شد بدست بیاورند، اما همچنان بدست‌آوردن چنین نتایجی امکان‌پذیر است. محققین غالباً دلایل خوبی برای استفاده از این نوع طرحها دارند. از جمله این دلایل می‌توان به این موارد اشاره کرد: نیاز به تخصیص منابع بیشتر به روایی سازه یا روایی بیرونی؛ الزامات کاربردی که بواسطه بودجه تحقیق، اخلاقیات یا مجریان طرح تحمیل می شود؛ یا محدودیتهای قانونی که هنگام ارزیابی یک درمان اجرا شده ایجاد می‌شود، هنگامی که شیوه ارزیابی پیش از اجرا طراحی نشده است. با توجه به اینگونه محدودیتها، بعضی اوقات اینگونه طرحهای شبه‌آزمایشی بهترین انتخاب هستند، اگرچه ممکن است استنباطهای علّی بدست‌آمده از این راه، ضعیفتر از استنباطهایی باشدکه از راههای دیگر بدست می‌آیند. در نتیجه، در این فصل ما به ارائه این طرحها و شرایطی که باعث می‌شود این نوع طرحها برای استنباطهای علّی توصیفی مفید باشند، خواهیم پرداخت. به علاوه، در این فصل به سه مورد موضوع دیگر نیز خواهیم پرداخت. اول، این طرحها به طور مکرر در تحقیقات میدانی بکار گرفته می‌شوند؛ برای مثال، در یکی از ارزیابیهای اخیر از برنامه ارزیابی آموزش پیش‌دبستانی، غالب مطالعات (۷۶٪) طرح پیش‌آزمون-پس‌آزمون تک‌گروهی استفاده کرده بودند، و غالب مطالعات باقی‌مانده همین طرح را بدون پیش آزمون مورد استفاده قرار داده بودند. برخی اوقات چنین کاربردهایی نشان‌دهنده یک باور اشتباه است، به این مضمون که عناصر طراحی مانند گروههای کنترل، یا پیش‌آزمونها غیرضروری یا غیرمطلوب هستند (حتی زمانی که استنباطهای علّی از درجه اولویت بسیار بالایی برخوردار است). در این فصل قصد داریم با نشان دادن هزینه‌هایی که این طرحها برای روایی درونی دربردارند، تردیدهایی را در مورد این باور قالبی مطرح نماییم. به این ترتیب محققین می‌توانند انتخاب کنند که آیا مایلند با وجود دیگر اولویتهای تحقیقاتشان، هزینه این طرحها را متحمل شوند؟ دوم، ما از این طرحها برای نشان دادن اینکه چطور تهدیدات مترتب بر روایی می‌توانند در مثالهای واقعی مسأله‌ساز باشند، استفاده می‌کنیم. این مسأله را از آن جهت مورد تأکید است که اهمیت داشتن نگاه نقادانه نسبت به این تهدیدها بیشتر از یادگرفتن مجموعه‌ای از طرحهای آزمایشی است. آشنایی درست با انواع مختلف این تهدیدها در این فصل و فصلهای آتی امکان تشخیص آنها را در مطالعات مختلف برای محقق تسهیل می‌کند. در نهایت، نگارندگان این طرحها را برای معرفی عناصر ساختاری مشترک تمامی طرحهای آزمایشی به کار می‌گیرند. این عناصر به محققین کمک می‌کند تا طرحهایی را بسازند که با توجه به شرایط خاص مطالعه آنها، روایی درونی قویتری را فراهم می‌آورد. این عناصر بارها و بارها در طرحهایی که در فصول آتی به آنها خواهیم پرداخت، مورد استفاده قرار می‌گیرند.

مختصری در باب منطق انجام شبه‌آزمایشها

طرحهای مورداشاره در فصل حاضر، شبه آزمایش هستند (آزمایشهایی که در آنها تخصیص تصادفی موارد به شرایط آزمون و کنترل انجام نمی شود، و بجز این مورد، اهداف و مشخصه‌های مشابهی با آزمایشهای تصادفی دارند). تحقیقات شبه‌آزمایشی برای سالهای متمادی مورد استفاده قرار گرفته‌اند. لیند (Lind, 1753) مقایسه شبه‌آزمایشی شش روش درمانی برای بیماری اسکوربوت را تشریح می‌کند. هارمن (Harman, 1936) اثر بروشورهای سیاسی منطق‌محور یا احساس‌محور را بر نتایج انتخابات در پنسیلوانیا مورد بررسی قرار داد. او سه سالن رأی‌گیری را که در آنها افراد بروشورهای حاوی اطلاعات احساس‌محور دریافت کرده بودند را با چهار سالن رأی‌گیری که در آنها، بروشورهای حاوی اطلاعات منطق‌محور توزیع شده بود جفت[۲۳۵] کرد. سالنها از نظر اندازه، تراکم جمعیت، ارزش ارزیابی‌شده زمین یا ملک، الگوهای پیشین رأی‌دهی، و موقعیت اقتصادی‌اجتماعی جفت شده بودند.

استنباط علّی در هر نوع از شبه‌آزمایشها باید با الزامات پایه‌ای روابط علّی تطابق داشته باشد. این الزامات عبارتند از: (۱) علّت باید قبل از اثر باشد، (۲) علّت باید با اثر همبستگی داشته باشد، و (۳) دلایل و تبیین‌های جایگزین برای رابطه علّی مورد نظر، منطقی و موجّه نباشند. آزمایشها و شبه‌آزمایشها هر دو مداخله یا دستکاری را به گونه‌ای اجرا می‌کنند که قبل از اثر اتفاق بیافتد. ارزیابی همبستگی و هم‌تغییری علّت و اثر نیز به سادگی در جریان انجام تحلیلهای آماری انجام می‌شود. در آزمایشها برای کنترل و بی‌اثر کردن تبیینهای جایگزین، باید اطمینان حاصل شود که متغیرهای موردنظر به صورت تصادفی در تمامی گروههای آزمایش و کنترل توزیع شده‌اند. از آنجاییکه در شبه‌آزمایشها تخصیص تصادفی وجود ندارد، در این طرحها بر اصول دیگری برای نشان دادن غیرموجّه بودن تبیین جایگزین تکیه می‌کنیم. در اینجا به سه مورد از این اصول در طرحهای شبه‌آزمایشی اشاره می‌نماییم.

- اصل اول، تشخیص و مطالعه تهدیدهای موّجه و معقول مترتب بر روایی درونی است. به مجرد اینکه این تهدیدها تشخیص داده شدند، احتمال اینکه این تهدیدها بتواند اثر مشاهده‌شده را تبیین نماید، قابل بررسی خواهد بود. در این فصل، مثالهای متعددی را برای نشان دادن اینکه چطور می‌توانیم با استفاده از این تهدیدها استنباطهای بدست‌آمده از شبه‌آزمایشها را نقد نماییم، ارایه خواهیم کرد.
- اصل دوم، اهمیت و اولویت کنترل کردن از طریق طرح آزمایش است. شبه‌آزمایشها از طریق اضافه کردن عناصر طراحی (مانند مشاهدات پیش‌آزمون بیشتر و گروههای کنترل بیشتر) سعی می‌کنند که یا از اثر متغیرهای مزاحم که اثری مشابه متغییر مستقل مورد مطالعه دارد جلوگیری کنند، و یا موجّه

235 Match

بودن اثر آنها را از طریق ارایه مستندات و شواهد تأیید نمایند. راه‌حل جایگزین برای کنترل از طریق طرح آزمایش، کنترلهای آماری هستند که سعی می‌کنند با استفاده از تعدیل آماری، اثرات مزاحم را از تخمین اثرات حذف نمایند. البته این کار بعد از این که مطالعه به پایان رسید انجام می‌شود. یقیناً کنترل از طریق طراحی و کنترل آماری همزمان باید به کار گرفته شوند. اما اولویت و ارجحیت با استفاده از کنترل از طریق طرح آزمایش است، و بهتر است کنترلهای آماری برای مواجهه با اثرات اندک باقی‌مانده، پس از بکارگیری کنترل از طریق طرح آزمایش، بکار گرفته شوند.

- سومین اصل برای کاهش اثرات موجّه متغیرهای مزاحم در شبه‌آزمایشها، *جفت‌کردن با الگوی همگن*[۲۳۶] است. در این روش، پیش‌بینی پیچیده‌ای درباره فرضیه علّی مورد نظر ارائه می‌شود، بطوریکه که کمتر تبیین جایگزینی بتواند با این الگوی پیچیده همخوان باشد. از جمله مثالهای آورده شده در این فصل در این خصوص، می‌توان به بکارگیری متغیرهای مستقل غیرهم‌ارز، و اثرات برهم‌کنشی[۲۳۷] پیش‌بینی‌نشده اشاره کرد. هرچه الگوی پیش‌بینی‌شده پیچیدگی بیشتری داشته باشد، احتمال کمتری وجود دارد که تبیین جایگزین بتواند الگویی مشابه آن تولید کند؛ و بنابراین با احتمال قویتری می‌توان گفت که مداخله موردنظر اثر واقعی داشته است.

هیچکدام از این سه اصل نمی‌توانند سهولت استنباط علّی یا منطق آماری متعالی همراه با تخصیص تصادفی را در اختیار ما بگذارند. در عوض، منطق استنباط علّی در شبه‌آزمایش نیازمند توجه دقیق، و با در نظر گرفتن جزئیات به شناسایی و کاهش توجیه‌پذیری تبیینهای علّی جایگزین است.

جدول ۴.۱: طرحهای شبه‌آزمایشی بدون گروه کنترل

طرح تک‌گروهی تنها پس‌آزمون X O_1
طرح تک‌گروهی تنها پس‌آزمون، با چندین پس‌آزمون مجزا و مستقل X_1 $\{O_{1A}\ O_{1B}\ldots O_{1N}\}$

236 Coherent pattern matching
237 Interaction

طرح تک‌گروهی پیش‌آزمون- پس‌آزمون $\boldsymbol{O_1}$ $\boldsymbol{X}$ $\boldsymbol{O_2}$
طرح تک‌گروهی پیش‌آزمون- پس‌آزمون با استفاده از پیش‌آزمون دوگانه O_1 O_2 X O_3
طرح تک‌گروهی پیش‌آزمون-پس‌آزمون با استفاده از یک متغیر وابسته غیرهم‌ارز $\{O_{1A.}\ O_{1B}\}$ X $\{O_{2A}\ O_{2B}\}$
طرح مداخله حذف شده O_1 X O_2 O_3 ⊼ O_4
طرح مداخله‌های مکرر O_1 X O_2 ⊼ O_3 X O_4

طرحهای بدون گروه کنترل

در این بخش، طرحهای بدون گروه کنترل را مورد بررسی قرار می‌دهیم (جدول ۴.۱). طرحهای بدون گروه کنترل تنها زمانی می‌توانند استنباطهای علّی قابل‌قبول ارایه کنند که بتوانند توجیه‌پذیر بودن تبیینهای جایگزین در مورد اثر مداخله را کاهش دهند. برخی طرحها در این زمینه ضعیف عمل می‌کنند، و برخی دیگر عملکرد بهتری دارند. با حرکت از طرحهای ضعیف به سمت طرحهای قویتر نشان خواهیم داد که چطور می‌توان طرحهایی ساخت که تهدیدهای بیشتر نسبت به روایی درونی را منتفی کنند.

طرح تک‌گروهی تنها با پس‌آزمون

در این طرح روی کسانی که مداخله را دریافت کرده‌اند یک پس‌آزمون انجام می‌دهیم. در این طرح نه گروه کنترل وجود دارد، و نه پیش‌آزمون. این طرح به شکل زیر نشان داده می‌شود. به گونه‌ای که X نشان‌دهنده مداخله و o_1 پس‌آزمون را نشان می‌دهد.

X O_1

چون پیش‌آزمونی وجود ندارد، دشوار می‌توان فهمید که آیا تغیری رخ داده است یا خیر. در ضمن نبودن گروه کنترل تشخیص اینکه در غیاب مداخله چه اتفاقی می‌افتد را سخت می‌کند. تقریباً تمامی تهدیدات مترتب بر

روایی درونی - به غیر از تقدم زمانی- در مورد این طرح مصداق دارند. برای مثال، خطر گذشت زمان تقریباً همواره وجود دارد، چون ممکن است دیگر رویدادها همزمان با اجرای مداخله رخ داده، و اثری مشابه به اثر مداخله ایجاد نموده باشند.

اگرچه، این طرح در موارد معدودی که دانش متنهابهی در مورد نحوه رفتار متغیر وابسته وجود دارد، می‌تواند سودمند باشد. برای مثال، دانش پایه محاسبات ریاضی در میان دانش‌آموزان دوره متوسطه در دبیرستانهای آمریکا به طور پایداری در سطح پایین قرار دارد. بنابراین اگر دانش‌آموزان پس از گذراندن یک دوره آموزش ریاضی بتوانند یک امتحان را با نمراتی بسیار بالاتر از حالتی که بتوان آن را شانسی دانست پشت سر بگذارند، می‌توان اثر مشاهده شده را به دوره آموزش ریاضی نسبت داد. دانش‌آموزان ریاضی چندانی در محیط خانه، از دوستان، تلویزیون، فعالیتهای فراغتی و یا حتی دیگر دوره‌های آموزشی یاد نمی‌گیرند. اما برای بدست‌آوردن استنباطهای علّی توصیفی واجد روایی، لازم است اثر مشاهده‌شده به اندازه‌ای بزرگ باشد که بوضوح شاخص باشد، و یا اینکه علّتهای جایگزین ممکن، شناخته شده و به روشنی غیرموجّه باشند، و یا اینکه هیچ تبیین جایگزین دیگری در زمینه مورد مطالعه عمل نکند (Campbell, 1975). این شرایط به ندرت در علوم اجتماعی مصداق دارد، و بنابراین این طرح به ندرت در این شکل ساده مفید واقع می‌شود.

ارتقاء طرح تک‌گروهی تنها پس‌آزمون، با استفاده از پس‌آزمونهای متعدد

طرحهای بدون پیش‌آزمون بیشتر برای مطالعات مبتنی بر نظریه که در حال آزمون یک نظریه هستند، بکار می‌روند. به این مطالعات، مطالعات تطبیق الگو[۲۳۸] (Campbell, 1966a; Trochim, 1985) یا تجانس (Rosenbaum, 1995a) گفته می‌شود. کشف قاتل بعد از وقوع جرم را به عنوان مثالی از یک چنین موقعیتی در نظر بگیرید. با توجه به اینکه کارآگاهان می توانند در کشف علل قتل موفق باشند، سیرون (Scriven, 1976) موفقیت کارآگاهان را وابسته به چند موضوع می‌داند: اهمیت و بزرگی اثر (جنازه)، در دسترس بودن الگویی از سرنخها که زمان و شیوه قتل را مشخص می‌کنند (پس‌آزمونهای متعدد)، و امکان و توانایی ربط دادن این سرنخها به شیوه‌های خاص ارتکاب جرم توسط مجرمینی (تبیین‌های جایگزین ممکن) که شیوه‌های خاصی را در ارتکاب جرم دارند، و این شیوه‌ها شبیه به شواهدی است که درجنایت مورد بررسی مشاهده شده است (یعنی الگوی رفتاری این مجرمین به جرم موردبررسی شبیه است). اگر بیش از یک مجرم الگویی مشابه الگوی جرم رخ داده دارد، این افراد را برای بازجویی، و گرفتن اثر انگشت آنها احضار می‌کنند (به این روش، روش امضاء گفته می‌شود).

238 Pattern matching

پاتولوژیستها با استفاده از این روش - با رویه‌ای شبیه کارآگاهان- شواهد جسد، موقعیت و زمان مرگ را برای بررسی علّت مرگ استفاده می‌کنند. آنها الگوی داده‌های موجود را با توضیحات موجود در متون علمی تطبیق می‌دهند، و دلایل ممکن برای مرگ را تشخیص می‌دهند. این سرنخها به مثابه پس‌آزمونهای مجزا، منحصربه‌فرد، و متعدد در این طرح شبه‌آزمایشی عمل می‌کنند.

$$X \qquad \{O_{1A} \;\; O_{1B} \;\; \dots \;\; O_{1N}\}$$

$\{O_{1A} \;\; O_{1B} \;\; \dots \;\; O_{1N}\}$ نشان‌دهنده مقیاسهای مرتبط با سازه‌های پس‌آزمونی متفاوت، ازA گرفته تا N، است که با الگوی اثرات بدست‌آمده از علّتهای شناخته‌شده (مثلاً مظنونین) تطبیق دارد. این با طرحی که در آن، تنها یک مقیاس مربوط به یک سازه در پس‌آزمون مورد بررسی قرار می‌گیرد (مثلا O_{1A}) و اساساً منطق تطبیق الگو در آن قابل استفاده نیست، متفاوت است.

اگرچه در این مثالها اثر معلوم بود، و به دنبال دلایل می‌گشتیم، اما در غالب اوقات در شبه‌آزمایشها عکس این موقعیت وجود دارد، یعنی علّت مشخص است (مثلاً دوره آموزش ریاضی جدید)، ولی اثر نامشخص بوده و به دنبال آن می‌گردیم (مثلاً چه تأثیری در میزان موفقیت دانش‌آموزان خواهد داشت). در این مواقع، منطق تطبیق الگو کمتر به کار می‌آید، زیرا علّت در اینجا اغلب یک نوآوری با یک الگوی ناشناخته از اثرات است. بنابراین اضافه کردن پس‌آزمونهای متعدد به طرح می‌تواند باعث افزایش خطای نوع اول شود، چون انسان به طور کلی تمایل دارد الگوهایی را حتی در داده‌های تصادفی پیدا کند (Fischhoff, 1975; Paulos, 1988; Wood, 1987).

بنابراین، تعیین و تشخیص الگوهای پشتیبانی‌کننده یک رابطه علّی، باید از قبل و با دقت انتخاب شود (این البته در زمینه‌هایی که دارای نظریه‌های مستحکم و قوی هستند امکان‌پذیر است). اما حتی در این حالت نیز لازم است که الگوی پیش‌بینی‌شده منحصربه‌فرد باشد. زیرا مثلاً اگر کتابهای آسیب‌شناسی برای سه بیماری تغییرات مشاهده‌شده را گزارش کرده باشند، نمی‌توانیم بر اساس این تغییرات میان آن سه بیماری تمایزی قائل شویم.

طرح تک‌گروهی پیش‌آزمون-پس‌آزمون

با اضافه کردن یک پیش‌آزمون (که متغیرهای وابسته را ارزیابی می‌کند) به طرح قبلی، به یک طرح پیش‌آزمون-پس‌آزمون تک‌گروهی می‌رسیم. در این طرح، یک پیش‌آزمون روی گروهی از پاسخگوها انجام می‌شود، سپس مداخله اجرا می‌شود و در نهایت یک پس‌آزمون با همان مقیاسها انجام می‌شود.

$$O_1 \qquad X \qquad O_2$$

اضافه کردن یک پیش‌آزمون اطلاعاتی ضعیف در مورد استنباطهای جایگزین مرتبط با زمانی که مداخله اتفاق نمی‌افتد، در اختیار ما قرار می‌دهد. اگرچه چون O_1 قبل از O_2 انجام می‌شود، این دو ممکن است به دلیلی غیر

از مداخله (مانند بلوغ، یا گذر زمان) با هم متفاوت باشند. برای مثال، جیسون، مک‌کوی، بلانکو و زولیک (Jason, McCoy, Blanco, & Zolik, 1981) اثر کمپینی را که از طریق توزیع جزوات آموزشی، و کیسه‌های پلاستیکی برای جمع‌آوری فضولات حیوانات خانگی، سعی در کاهش ریخته شدن فضولات سگ در خیابانهای شیکاگو داشت، مورد بررسی قرار دادند. پس از کمپین میزان فضولات حیوانات به میزان قابل‌توجهی کاهش یافته بود، اما مشخص نبود که در زمان انجام کمپین اتفاق دیگری نیافتاده که اثری مشابه داشته باشد (مثلاً اگر آن روزها اتفاقاً بارانی باشد، و مردم سگهایشان را کمتر بیرون آورده باشند).

این طرح را می‌توان روی همان افراد، یا افراد متفاوتی که هر دو پیش‌آزمون و پس‌آزمون را دریافت می‌کنند اجرا کرد. هنگامی که روی همان افراد انجام می‌شود، طرح درون‌گروهی[۲۳۹] گفته می‌شود[۲۴۰]. دوکارت (Duckart, 1998) طرح درون‌گروهی پیش‌آزمون-پس‌آزمون را برای ارزیابی اثر یک برنامه روی کاهش سرب محیطی در خانوارهای شهری کم درآمد در بالتیمور مورد استفاده قرار داد. برنامه برای کاهش مقدار سرب در هر خانه از دو روش ایجاد تغییرات فیزیکی و آموزش دادن استفاده می‌کرد. با استفاده از دستمالهای تشخیص سرب، میزان سرب در مکانهای مختلف هر خانه در سه مرحله ۱) پیش‌آزمون، ۲) بلافاصله پس از مداخله (پس‌آزمون)، ۳) شش ماه بعد (مطالعه پیگیری) محاسبه شد. کاهش معناداری در سطح سرب پیش و پس از آزمون مشاهده شد، و این کاهش (البته به میزان کمتری) در پس‌آزمون شش ماه بعد نیز مشاهده شد. از آنجایی که سطح سرب در بازه‌های زمانی کوتاه ثابت است، کمتر احتمال دارد بتوان تغییراتی را با یک گروه کنترل پیدا کرد. بنابراین در مورد این موضوع، طرح پیش‌آزمون-پس‌آزمون بدون گروه کنترل طرح مناسبی است.

اگرچه دوکارت (Duckart, 1998) اشاراتی به تهدیدهایی روایی در این طرح دارد. از نظر روایی درونی، بلوغ[۲۴۱] می‌تواند تغییرات سطح سرب را در فاصله پیش‌آزمون تا آزمون پیگیری[۲۴۲] تحت‌تأثیر قرار دهد، زیرا سطح غبار سرب در تابستان بیشتر از زمستان است، و بیشتر آزمونهای پیگیری در زمستان صورت گرفته بود. گذر زمان نیز می‌توانست یک تهدید محسوب شود، زیرا همزمان با انجام این مطالعه، یک شرکت دیگر نیز در بالتیمور خدماتی ارایه می‌کرد که می‌توانست سطح سرب موجود در بعضی خانه‌های درون نمونه را تحت تأثیر قرار دهد. انجام آزمون نیز می‌تواند نوعی تهدید محسوب شود، زیرا بازخوردهای پیش‌آزمون به شهروندان در مورد سطح سرب

239 within-subject

۲۴۰ در اینجا عامل درون–گروهی زمان است. به این معنی که اندازهی گیری های مکرر در پیش‌آزمون و پس‌آزمون درون هر واحد (فرد). در دیگر طرحهای درون–گروهی بیش از یک شرایط را می توان روی افراد یکسان اجرا کرد. در مقابل، آزمایشی که چندین شرایط، و واحدهای متفاوت در هریک از شرایط دارد، طرح میان–گروهی گفته می شود (S.E. Maxwell & Delaney, 1990). طرح های درون–گروهی از طریق کنترل تفاوتهای فردی میان افراد درون هر شرایط، می توانند توان آماری را افزایش دهند؛ و بنابراین این طرحها (در قیاس با طرحهای بین–گروهی) می توانند افراد کمتری (نمونه کوچکتری) را برای آزمون همان تعداد مداخله داشته باشند. اگرچه طرحهای درون–گروهی می توانند زمینه سازی بروز اثرات خستگی، تمرین، سرریز و ترتیب شود. برای جلوگیری از اختلاط اثر اینگونه متغیرهای مزاحم، یا ترتیب ارائه مداخلهها در طرح درون–گروهی برای هر فرد تصادفی سازی می شود و یا مداخله ها با دقت متوازن–معکوس می شوند به این معنی که برخی افراد مداخله ها را با ترتیب اول (یعنی اول A بعد B) می گیرند و برخی دیگر با ترتیب دوم (اول B و بعد A).

241 maturation

242 Follow-up

می‌توانسته باعث شده باشد آنها با دقت و وسواس بیشتری نظافت کرده باشند، که این خود می‌تواند سطح سرب را کاهش دهد، حتی اگر مداخلات انجام شده غیراثربخش بوده باشند. در خصوص ریزش، در حدود یک سوم شرکت‌کنندگان مطالعه را پس از پیش‌آزمون ترک کردند. این شهروندان ممکن است افرادی با کمترین میزان انگیزه و یا سطح همکاری بوده باشند. روایی نتایج آماری ممکن است بواسطه اندازه کوچک نمونه و نتیجتاً توان آماری اندک در آزمونهایی که معنادار نبودند کاهش یابد، و توان می‌توانسته حتی بیش از این نیز کاهش یابد، زیرا اجرای مداخله بطور قابل‌توجهی از یک مکان به مکان دیگر متفاوت بوده است. روایی سازه نتایج اندازه‌گیری نیز ممکن است کاهش یافته باشد، زیرا همان فردی که مداخلات را در خانه‌ها اجرا کرده بود، خود نتایج را نیز اندازه‌گیری کرده بود. این افراد ممکن است اندازه‌گیری را به‌گونه‌ای انجام داده باشند، که نتایج اجرای مداخلات مطلوب به نظر برسد.

این مثال بطور اخص برای نشان دادن منطق لحاظ کردن و در نظر گرفتن تهدیدات روایی در مورد استنباطهای علّی مفید است. دوکارت (Duckart, 1998) دو کار عمده انجام داد. اول، وی نشان داد که تهدیدات نه تنها ممکن، که موجّه هستند. او با ارایه داده‌هایی در رابطه با تهدیداتی که در مطالعه وی امکان ایجاد مشکل داشتند، این مسأله را نمایش داد. تحقیقات پیشین نشان می‌دادند که سطوح غبار سرب در زمستان کمتر از تابستان هستند. دوم، او نشان داد که این تهدید اثری شبیه اثر مورد انتظار وی از متغیر علّت دارد (به این معنی که او میزان کمتری از سرب را پس از اجرای مداخله بدست آورد، که نتایجی شبیه به این در زمستان نیز دیده می‌شود). بنابراین، این سطح پایین‌تر می‌توانسته به دلیل اثر فصل بوده باشد. هر دو این شرایط این استنباط را که اجرای برنامه موجب بروز نتایج مشاهده شده بوده را تضعیف می‌کنند. برای مثال تصور کنید که نتایج تحقیقات پیشین نشان می‌داد که سطح غبار سرب در تابستان کمتر از زمستان است. این نتایج نمی‌توانست سطح پایین مشاهده شده در مطالعه دکارت را تبیین کند، و بنابراین نمی‌توانست تهدیدکننده روایی درونی باشد. برای این که تهدیدی موجّه باشد، و بتواند روایی را به خطر بیاندازد، باید اولاً، به اندازه کافی بزرگ باشد تا بتواند اثر قابل‌توجه مشاهده‌شده را توضیح دهد. ثانیاً، هم‌جهت با اثر علّت مورد بررسی باشد، و نه در جهت خلاف آن. مثلاً در این مثال، اگر میزان سرب به طور مثال در تابستان کمتر از زمستان بود، این مسأله تهدیدی برای روایی مطالعه به حساب نمی‌آمد.

در مجموع پژوهشگران علوم اجتماعی، کمتر می‌توانند با استفاده از طرحهای پیش‌آزمون-پس‌آزمون در مطالعات میدانی، روابط متقن علّی را تأیید کنند؛ مگر اینکه نتایج با احتیاط مورد استفاده قرار بگیرند، و یا اینکه بازه زمانی میان پیش‌آزمون و پس‌آزمون کوتاه باشد. افرادی که این طرحها را در نظر می‌گیرند، باید عناصر بیشتری را به طرح اضافه کنند.

ارتقاء طرح پیش‌آزمون-پس‌آزمون تک‌گروهی با اضافه کردن پیش‌آزمون دوگانه[۲۴۳] **(دو پیش‌آزمون)**
توجیه‌پذیر بودن تهدیدات ناشی از بلوغ و رگرسیون را می‌توان از طریق اضافه کردن یک پیش‌آزمون دوم به طرح آزمایش (و انجام آن قبل از پیش آزمون اول) انجام داد.

$$O_1 \qquad O_2 \qquad X \qquad O_3$$

دو پیش‌آزمون به عنوان یک «اجرای پایه[۲۴۴]» عمل کرده، و سوگیریهایی که ممکن است در تخمین اثر مداخلهها طی مرحله O_2 به O_3 وجود داشته باشد را روشن می‌سازند. برای مثال، مارین و همکارانش (Marin et al., 1990) در مطالعه‌ای اثر یک کمپین ترک سیگار از طریق اطلاع‌رسانی متناسب با فرهنگ اسپانیایی-تبارهای آمریکایی را مورد بررسی قرار دادند. پیش‌آزمون اول در ۱۹۸۶ انجام شد. پیش‌آزمون دوم در تابستان ۱۹۸۷، و پس‌آزمون در ۱۹۸۸ صورت گرفت. نتایج نشان داد که سطح اطلاعات پس از کمپین (یعنی در زمان O_3) بسیار بیشتر از O_2 و همچنین بسیار بیشتر از روند بلوغ میان O_2 و O_1 بوده است. البته اگر تعداد بیشتری پیش‌آزمون می‌داشتیم، می‌توانستیم رابطه غیرخطی در تغییرات رخ داده بواسطه بلوغ در پیش‌آزمون را شناسایی کنیم.

ارتقاء طرح پیش‌آزمون-پس‌آزمون با استفاده از یک متغیر وابسته غیرهم‌ارز[۲۴۵]
اضافه کردن یک متغیر وابسته غیرهم‌ارز به طرح آزمایش به صورت زیر نشان داده می‌شود. در این طرح A و B نشان‌دهنده مقیاسهای متفاوتی است که از یک گروه واحد در دو زمان ۱ و ۲ پرسیده می‌شود.

$$\{O_{1A} \text{ و } O_{1B}\} \qquad \text{X} \qquad \{O_{2A} \text{ و } O_{2B}\}$$

مقیاسهای A و B سازه‌های مشابهی را اندازه گیری می‌کنند. انتظار می‌رود مقیاس A (برون‌داد) به دلیل اعمال مداخله تغییر کند. در حالی که مقیاس B (متغیر وابسته غیرهم‌ارز) به دلیل مداخله تغییر نمی‌کند. بلکه انتظار می‌رود این مقیاس به تهدیدات مهم مترتب بر روایی درونی عکس‌العمل نشان دهد. البته پاسخی که شبیه تغییرات A است [توضیح مترجم: یعنی اگر تهدیداتی نسبت به روایی درونی وجود داشته باشد، و تبیین جایگزینی برای توضیح موردبررسی وجود داشته باشد، متغیر B به دلیل آن تبیین جایگزین، تغییر می‌کند]. برای مثال رابرتسون و روسیتر (Robertson & Rossiter, 1976) برای مطالعه ترجیحات کودکان در مورد اسباب‌بازیهای تبلیغ‌شده در طول مدت‌زمان کریسمس، متغیر وابسته غیرهم‌ارزی را بکار گرفتند. مطالعه آنها نشان داد که سطح مطلوبیت (متغیر وابسته) اسباب‌بازیهای تبلیغ شده در طول این مدت (یعنی نوامبر و دسامبر)، بیش از موارد تبلیغ‌نشده

243 A double pretest
244 Dry run
245 Non-equivalent

(متغیر غیرهم‌ارز وابسته) افزایش یافته بود. این طرح موجّه بودن بسیاری از تهدیدهای مترتب بر روایی درونی را کاهش می‌داد. مثلاً با توجه به پرداختن مکرر افراد به موضوع خرید هدیه و اسباب بازی در طول کریسمس در فرهنگ آمریکایی، تهدید گذر زمان در این مطالعه یک تهدید موجّه است. اگرچه این متغیر (گذر زمان)، تمامی انواع اسباب‌بازی (هم تبلیغ‌شده و هم تبلیغ‌نشده) را تحت تأثیر قرار می‌دهد. البته اگر اسباب‌بازیهای تبلیغ‌شده در تلویزیون، در رادیو و روزنامه هم تبلیغ شوند، یک مشکل دیگر ناشی از گذر زمان بوجود می‌آید. چه کسی می‌تواند بگوید که اثر مشاهده شده ناشی از تبلیغ در تلویزیون بوده است؟ یا رگرسیون آماری چطور می‌تواند اثرگذار باشد اگر تولیدکنندگان اسباب‌بازیهایی را تبلیغ کرده باشند که به طور معمول فروش کمتری در طول این دوره داشته‌اند؟

مک کیلیپ و همکارانش (McKillip & Baldwin, 1990) دریافتند که دادن بازخورد درباره دزدی چیپس سیب‌زمینی از قفسه چیپس، میزان این نوع دزدی را کم می‌کند، اما تأثیری بر میزان دزدی بستنی، شیر و ساندویچ ندارد. این طرح می‌تواند برای دامنه وسیعی از مطالعات مفید بوده و اغلب می‌تواند ارائه‌دهنده استباطهای علّی باشد، البته اگر هر دو متغیرهای مستقل (مورد بررسی و غیرهم‌ارز) به میزان مساوی در معرض عوامل محیطی (یکسان) قرار بگیرند.

طرح مداخله حذف شده

این طرح یک پس‌آزمون سوم را به طرح پیش‌آزمون-پس‌آزمون (O_2 و O_1) اضافه می‌کند، و سپس مداخله را (✕ نشان‌دهنده حذف مداخله است) قبل از آن که اندازه‌گیری نهایی(O_4) انجام شود، حذف می‌کند.

$$O_1 \quad X \quad O_2 \qquad\qquad O_3 \quad \times \quad O_4$$

هدف آن است که نشان دهیم متغیر نتیجه‌ای با وجود و عدم‌وجود مداخله، بالا و پایین می‌شود؛ نتایجی که اگر ناشی از مداخله نباشند، تنها می‌توانند در اثر وجود یک عامل تهدیدکننده روایی، و به شکلی مشابه در طول همان بازه زمانی بوجود بیایند. حرکت از O_1 به O_2 یک فاز آزمایشی است، و حرکت از O_3 به O_4 فازی دیگر. فرضیه مورد بررسی در بخش دوم دقیقاً متضاد فرضیه بخش اول است. اگر قسمت اول یعنی از O_1 به O_2 پیش‌بینی‌کننده یک افزایش باشد (یعنی با اعمال مداخله)، قسمت دوم و در حرکت از O_3 به O_4 (یعنی در عدم‌وجود مداخله) پیش‌بینی‌کننده کاهش است. این الگوی رفتاری متغیر تنها زمانی می‌تواند رخ دهد که اثرات مداخله با حذف مداخله طی مدت‌کوتاهی از بین برود [بنابراین این طرح آزمایشی برای مداخلههایی که اثر ماندگار دارند مناسب نیست]. حتی اثری که بخشی از آن در فاصله O_2 به O_3 باقی مانده باشد می‌تواند در خلاف

جهت اثر معکوس مورد انتظار در فاصله O_3 به O_4 سوگیری ایجاد کند. شکل ۴.۱ قابل تفسیرترین نتایج حاصل از طرح مداخله حذف‌شده را به تصویر می‌کشد.

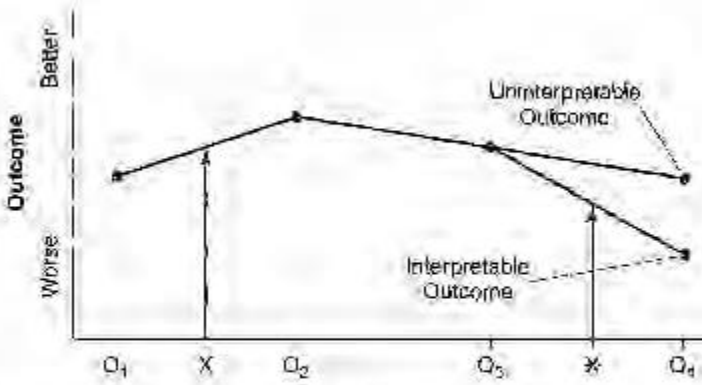


FIGURE 4.1 Generally interpretable outcome of the removed-treatment design

تأمین روایی آماری در این طرح یک مشکل است، چون الگوی نتایج میتواند حتی بواسطه یک داده پرت[۲۴۶] مورد تحریف قرار گیرد. بنابراین نمونه‌های بزرگ و مقیاسهای قابل‌اتکاء از ملزومات استفاده از این طرح است. جدای از این، حذف بعضی مداخله‌ها ممکن است غیراخلاقی باشد یا منجر به خستگی و فرسودگی شود، که این خود می‌تواند با مقیاسهای خشونت، رضایتمندی یا عملکرد همبستگی داشته باشد، و اثری مشابه رنجیده خاطر شدن گروه کنترل[۲۴۷] یا احساس مورد بی‌عدالتی واقع شدن، و در نتیجه اتخاذ اقدامات جبرانی شبیه مداخله توسط خود افراد[۲۴۸] شود. در این حالت، بهتر است از این طرح استفاده نکنید.

این طرح هنگامی که شرکت‌کنندگان خودشان دریافت مداخله را متوقف می‌کنند نیز می‌تواند به طور طبیعی اتفاق بیافتد، البته هنگامی که علّت ترک مطالعه چیزی بی‌ارتباط با مداخله باشد. اگرچه این ندرتاً اتفاق می‌افتد؛ بنابراین هنگامی که پاسخ‌دهندگان خودشان تصمیم به توقف مداخله می‌گیرند، باید با احتیاط بسیار حرکت کرد. برای روشن شدن این موضوع، مطالعه‌ای را در نظر بگیرید که اثرات مترتب بر نگرش ورود به نقش کاری جدید را بررسی می‌کند. برای مثال، فرض کنید فردی سمت شغلی‌اش ارتقاء پیدا می‌کند و سرپرست می‌شود (X)؛ در بازه زمانی O_1 به O_2 نگرش موافق مدیریت (یا همراهی با مدیران) در وی افزایش می‌یابد، اما شکل جدید ارتباط با مدیران موردپسند وی واقع نمی‌شود. تا زمان O_3 ، نگرش موافق مدیریت در وی کاهش می‌یابد. این فرد می‌تواند از این سمت استعفا دهد، یا این که این سمت از وی پس گرفته شود در نتیجه سطح نگرش از

246 Outlier
247 Resentful demoralization
248 Compensatory rivalry

O_3 به O_4 بیشتر کاهش می‌یابد (در مقایسه با O_1 به O_2). در این حالت محقق باید تصمیم بگیرد که آیا تغییر رخ داده طی O_3 به O_4 در نتیجه تغییر شغلی اتفاق افتاده، و یا دنباله طبیعی O_3 بوده است. اگر O_3 به O_4 از نظر اندازه شبیه به اندازه تفاضل O_2 به O_3 باشد، دومی محتمل‌تر است (نمودار ۴.۱ را نگاه کنید). در حالی که شیب تندتر کاهش نگرش از O_3 به O_4 نسبت به O_2 به O_3 نشان‌دهنده آن است که ورود به شغل یا موقعیت شغلی خاص موجب افزایش نگرش موافق مدیریت بوده است.

لیبرمن (Liberman, 1956) شکل ساده‌تری از این طرح را بکار گرفت. وی نمونه‌ای از مردان قبل از اینکه سرپرست شوند، بعد از اینکه سرپرست شوند، و در نهایت، بعد از آنکه به موقعیت شغلی کارگری بازگشتند، انتخاب کرد. تنها سه دور اندازه‌گیری اجرا شد. از این رو تفاوت میان O_1، O_2 و O_3 می‌تواند به دو دلیل اتفاق افتاده باشد. ۱) می‌تواند در اثر تغییر نقش باشد که باعث تغییر نگرش شده است؛ ۲) می‌تواند به دلیل تنزل مقام سرپرستی باشد که نگرشهای وی کمتر با مدیران بالادستی هماهنگ بوده است. اضافه کردن دور چهارم مشاهده (اندازه گیری) کمک می‌کند تا این احتمالات را بررسی کنیم. مشاهدات باید در فاصله های زمانی مساوی انجام شوند تا امکان ارزیابی تغییرات خطی را میسر نماید. اگر فاصله زمانی میان O_3 و O_4 طولانی‌تر از فاصله زمانی میان O_2 تا O_3 باشد، مقایسه تفاوت میان O_2 و O_3 با تفاوت میان O_3 و O_4 چندان معنادار نخواهد بود.

طرح مداخله مکرر

ممکن است در مواقعی ابتدا مداخله را اعمال نموده، حذف نماییم، و سپس مجدداً اعمال نماییم تا در طی زمان بررسی کنیم چطور مداخله و نتایج در طول زمان کوواریانس دارند:

$$O_1 \quad X \quad O_2 \quad \asymp \quad O_3 \quad X \quad O_4$$

یک قدم فراتر از طرح حذف مداخله، در طرح مداخله مکرر فرض بر این است که تهدیدات روایی معدودی وجود دارند که بتوانند اثری شبیه رابطه نزدیک میان اعمال و حذف مداخله از یک سو، و تغییرات موازی نتایج از سوی دیگر (با همان کیفیت) را ایجاد کنند. این تهدیدات باید دقیقاً با همان شیوه‌ای که مداخله اعمال و حذف می‌شود، بیایند و بروند. چیزی که عموماً غیرمحتمل است. اگرچه در صورتی که اثر مداخله گذرا نباشد، دشوار است که بتوان از طریق حذف مداخله جهت اثر را معکوس نمود؛ یا اگر مداخله تا مرحله O_3 اثر سقف ایجاد نماید، این موجب می‌شود تا اعمال دوباره مداخله نتواند اثر مشخصی داشته باشد. بهترین و قابل تفسیرترین نتایج در این طرح وقتی است که O_1 نسبت به O_2 تفاوت کند و O_2 نسبت به O_3 در جهت عکس اثر قبلی تفاوت کند، و O_3 نسبت به O_4 با حالتی شبیه تفاوت O_1 نسبت به O_2 تغییر کند. پاورز و انجلین (Power & Anglin, 1993) این طرح را برای مطالعه اثر متادون روی میزان مصرف معتادین بکار گرفت. طی زمانی که معتادین

متادون مصرف می‌کردند، مصرف موادمخدر در آنها به میزان قابل‌توجهی کاهش یافت، و سپس با قطع تجویز متادون، دوباره مصرف موادمخدر آنها افزایش یافت، و دوباره با تجویز مجدد متادون مصرف موادمخدر در میان معتادین کاهش یافت. این طرح به طور مکرر از سوی روانشناسان رفتاری بکار گرفته می‌شود. به نظر می‌رسد علّت رواج گسترده این طرح در میان این محققین آن است که این طرح مشتمل بر حداقل یکبار تکرار اثر مداخله است، و از این رو، این تکرار تأمین‌کننده یکی از اصول اولیه تحقیق باکیفیت، یعنی قابلیت بازتولید[۲۴۹] است.

از جمله تهدیدات نسبت به روایی در این طرح، احتمال رخ دادن بلوغ است. مثلاً اگر O_4 و O_2 در سه‌شنبه صبح اندازه‌گیری شده باشند، در حالی که O_3و O_1 در جمعه عصر، تفاوت مشاهده شده در بهره‌وری می‌تواند بیشتر به دلیل تفاوتهای عملکرد در روزهای مختلف هفته باشد، تا به دلیل مداخله (با توجه به این که میزان بهره‌وری در ابتدا، و انتهای هفته متفاوت است). همچنین باید رویدادهای خاصی که طی بازه زمانی اعمال و حذف مداخله اتفاق افتاده، و می‌توانسته اثری مشابه مداخله موردنظر داشته باشد را در نظر بگیریم. در مجموع، این طرح از نظر روایی درونی طرح مستحکمی است؛ علی‌الخصوص زمانی که محقق اعمال و حذف مداخله را کنترل می‌کند. درمقابل، این طرح می‌تواند از نظر روایی بیرونی و روایی نتایج آماری آسیب‌پذیر باشد. برای مثال، بسیاری از نمودارهای عملکرد در مطالعات هاثورن (Roethlisberger & Dickson, 1939) در مورد زنان کارگر هستند که تعداد آنها شش نفر است (که رقم اندکیست)، و این افراد عکس‌العملهای بسیار متفاوتی نسبت به مداخله از خود نشان دادند؛ بنابراین نمی‌توانیم با اطمینان دریابیم که نتایج تا چه اندازه از تأثیر خطای نمونه‌گیری مصون بوده است. یقیناً بکارگیری نمونه‌های بزرگ و آزمونهای آماری قویتر می‌تواند در رفع این کاستی‌ها مفید واقع شود. همچنین در این طرح ها، زمانی که شرکت‌کنندگان متوجه اعمال و حذف مداخله و اعمال مجدد آن می‌شوند، روایی سازه به خطر می‌افتد. شرکت‌کنندگان ممکن است حدسیاتی در مورد چرایی این تغییرات در مداخله بزنند، و نسبت به آن عکس‌العمل نشان دهند. از سویی دیگر، احتمال دارد افراد مورد آزمایش در این طرحها دچار این احساس شوند که [بواسطه حذف مداخله در فاصله زمانی میان O_2 به O_3] مورد بی‌عدالتی واقع شده‌اند، و در نتیجه بخواهند خود اقداماتی جبرانی شبیه آنچه درمداخله انجام می‌شود اتخاذ نمایند[۲۵۰] ؛ که این می‌تواند مسأله‌ساز باشد. این مشکل می‌تواند داده‌های بدست‌آمده در زمان O_3 را تحت تاثیر قرار داده، و تفسیر افزایش مشاهده‌شده میان O_3 و O_4 (یعنی زمانی که مداخله مجدداً اعمال می‌شود) را هرچه بیشتر دشوار سازد.

249 reproductivity
250 Resentful demoralization

بنابراین، این طرح برای مداخله‌هایی مناسب است که ۱) اثرات گذرا داشته باشند، ۲) جلب توجه زیادی نکنند، ۳) فاصله زمانی طولانی میان مداخله اول و اجرای مجدد آن داشته باشد، و ۴) در آنها متغیرهای مداخله‌گری که سیکلهای زمانی تداخل‌کننده با زمان اجرای اول، حذف و اجرای مجدد مداخله دارند، وجود نداشته باشند. همچنین این طرحها زمانی اثربخش‌تر هستند، که اجرای مجدد مداخله مکرر بوده، و بطور تصادفی در طول زمان توزیع شود، در نتیجه نوعی آزمایش تصادفی ایجاد شود که در آن بلوکهای زمانی واحدهای تخصیص را شکل می‌دهند (Edington, 1987, 1992). البته نمی‌توان در همه پروژه‌های تحقیقاتی الزاماً همه این شرایط را تأمین نمود.

اجرای طرح مداخله مکرر با استفاده از پیش‌آزمون پس‌نگر[۲۵۱]

پاورز و آنجلین (Power & Anglin, 1993) این طرح را برای ارزیابی پس‌نگر اثرات متادن بر معتادین به هرویین که سابقه شرکت در طرحهای درمانی متفاوتی درگذشته داشتند، اجرا کردند. آنها از معتادین خواستند تا سابقه مصرف موادمخدر و تاریخ درمان خود را به صورت پس‌نگر یادداشت کنند. این دو محقق پس از بررسی این سرگذشتها دریافتند که متادن مصرف مواد مخدر را در طول دوره درمان با متادن کاهش می‌دهد، اما در فاصله میان درمانها اثری ندارد. متأسفانه، قضاوتهای پس‌نگر می‌تواند تا حدزیادی دچار سوگیری باشد (Silka, 1989). برای مثال، این تمایل در افراد وجود دارد که تخمینهای بزرگتری در مورد حوادث بد زمان حال در مقایسه با حوادث بد رخ داده در گذشته داشته باشند (به این معنی که حوادث بد در گذشته را کمتر به خاطر می‌آورند). بنابراین نتایج حاصل از خود-اظهاری[۲۵۲] پس‌نگر با نتایج حاصل از خود-اظهاری آینده‌نگر تفاوت دارد (Widon, Weiler, & Cottler, 1999; G.S. Howard et al., 1979). عوامل مختلفی می‌توانند پس‌آزمون پس‌نگر را تحت تأثیر قرار دهند. مثل اینکه آیا موضوع مورد تخمین به راحتی قابل مخدوش شدن است یا خیر (مثل شناختها در مقایسه با رفتارها)، موضوعی که باید به خاطر آورده شود مربوط به چه مدت قبل است، مشخصات تقاضا چیست (مثلاً پاسخهای مخدوش‌شده مربوط به رفتارهای غیرقانونی)، اصلاحات جزئی مورد نیاز است یا کلی (رویدادهای خاص را کمتر می‌توان به طور دقیق به خاطر آورد)، چه احساساتی بواسطه یادآوری اطلاعات برانگیخته می‌شود (برای مثال یادآوری یک درد یا واقعه دردناک) (به خاطر آوردن یک درد، Babock, 1998). پاره‌ای اوقات پیش‌آزمونهای پس‌نگر می‌تواند با استفاده از دیگر منابع صحت‌سنجی[۲۵۳] شود. برای مثال، پاورز و آنجلین (Power & Anglin, 1993) تاریخهای درمان گزارش‌شده توسط افراد را با سوابق ذخیره شده موجود در درمانگاه مقایسه و صحت‌سنجی کردند. اگرچه غالباً میزان دانسته‌های موجود مبتنی بر شواهد تجربی در زمینه هر طرح

251 Retrospective
252 Self-report
253 Cross-validated

آزمایشی ناچیز است. بنابراین لازم است تا بار دیگر به دیدگاه کمپبل و استنلی (Campbell & Stanley, 1963) اشاره کنیم که «با در نظر گرفتن عوامل اُتیستی مخدوش‌کننده حافظه، و در نتیجه مخدوش‌کننده گزارشهای یادآوری وقایع، اینگونه داده ها (داده‌های پیش‌آزمون پس‌نگر) را به هیچ عنوان نمی‌توان جدی گرفت» (ص ۶۶). پس‌آزمونهای پس‌نگر باید به‌صورت مکملی برای دیگر عناصر در نظر گرفته شوند. نباید این داده‌ها را به خودی‌خود مورد استفاده قرار داد بلکه باید به عنوان آخرین روش و با احتیاط به کار گرفته شوند.

مطالعه پاورز و آنجلین (Power & Anglin, 1993) دچار مشکل ریزش شرکت‌کنندگان در طول آزمایش نیز بود. آنها می‌بایست تنها معتادینی را بررسی می‌کردند که برای درمان باز می‌گشتند. بنابراین اندازه نمونه آنها با حرکت از دوره درمان اول به دوم، سوم و چهارم به تدریج کوچک و کوچکتر می‌شد. بیمارانی به درمان باز می‌گشتند که نتوانسته بودند پاک بمانند، و دیگران به این دلیل بازنگشته بودند که توانسته بودند از شر هروئین خلاص شوند. بنابراین، اگر کل نمونه تا پایان مطالعه دنبال می‌شدند، این نتیجه‌گیری کلی که درمان با متادن منجر به نوعی موفقیت موقت می‌شود، زیر سوال می‌رفت. این مسأله، یک مشکل عمومی برای مطلعات طولانی‌مدت[۲۵۴] (یا طولی) است، که شرکت‌کنندگان آن برای مدت‌طولانی دنبال می‌شوند. افرادی که در طول مطالعه از ادامه آن منصرف می‌شوند، به طور سیستماتیک با افرادی که در مطالعه می‌مانند تفاوت دارند. بسیاری از مشکلات مطرح شده در این بخش را می‌توان تنها با اضافه کردن یک گروه کنترل مستقل مرتفع کرد. برای مثال، در مطالعه پاورز و انجلین (Power & Anglin, 1993) افزودن معتادینی که تحت درمان متادن نبودند می‌توانست زمانبندی و فراوانی تغییرات در گزارشهای پیش‌آزمون پس‌نگر را در مراجعینی که مداخله (متادن) را دریافت نکرده‌اند روشن سازد. این می‌تواند نشان‌دهنده اثر ریزش در طول زمان باشد. دو نوع گروه کنترل می‌توانست مورد استفاده قرار بگیرد. یکی گروهی که لازم است تا برای ارزیابی به کلینیک برگردند؛ در این گروه، به میزان گروه آزمون، ریزش وجود خواهد داشت. دوم، گروهی که به صورت فعال توسط جامعه اطرافیان مورد پایش و مراقبت واقع می‌شود؛ در این گروه ریزش اتفاق نخواهد افتاد. در اینجا به بحث و مداقه بیشتر در مورد این نوع گروههای کنترل خواهیم پرداخت، و بحث را به شبه‌آزمایشهایی که دارای گروه کنترل هستند، اما پیش‌آزمون ندارند محدود می‌کنیم.

طرحهای شبه‌آزمایشی که گروه کنترل دارند، اما پیش‌آزمون ندارند

یک روش کلاسیک برای نشان دادن یک استنباط نقیض یا مقابل[۲۵۵] این است که یک گروه کنترل که هیچ نوع مداخله دریافت نمی‌کند، را به مطالعه اضافه کنیم. گروه کنترل باید به شیوه‌ای انتخاب شود که تا بیشترین حد ممکن به گروه آزمون شباهت داشته باشد (D'Agostino & Kwan, 1995). از نظر شیوه نمایش طرح، خط

254 longitudinal
255 counterfactual

نقطه چین (-----) بین گروهها نشان می‌دهد که گروهها به صورت تصادفی انتخاب نشده‌اند. ضمناً این نوع گروهها را با حروف انگلیسیNR نشان می‌دهند که این حروف به تخصیص غیرتصادفی[۲۵۶]اشاره دارد. جدول ۴.۲ طرحهای شبه‌آزمایشی که گروههای کنترل دارند، اما در آنها هیچ اندازه‌گیری پیش‌آزمون روی متغیر برون‌داد صورت نمی‌گیرد را نشان می‌دهد.

جدول ۴.۲ طرحهای شبه‌آزمایشی که گروههای کنترل دارند، اما پیش‌آزمون ندارند

طرح پس‌آزمون با گروه‌های غیرهم‌ارز

NR	X	O_1
NR		O_2

طرح پس‌آزمون با استفاده از یک نمونه پیش‌آزمون مستقل

NR	O_1	X	O_2
NR	O_1		O_2

طرح پس‌آزمون با استفاده از پیش‌آزمون نماینده [۲۵۷]

NR	O_{A1}	X	O_{B2}
NR	O_{A1}		O_{B2}

طرح تنها پس‌آزمون، با گروههای غیرهم‌ارز

در طرح پس‌آزمون با گروههای غیرهم‌ارز، یک گروه کنترل به طرح تک‌گروهی پس‌آزمون اضافه می‌کنیم. این طرح را هنگامی می‌توان به کار گرفت که اجرای یک مداخله قبل از این که با محقق هماهنگ شده باشد، آغاز شده است. بنابراین در این طرح مشاهدات و اندازه‌گیریهای پیش‌آزمون (با همان مقیاسی که اندازه‌گیریهای پس‌آزمون انجام می‌شود) در دسترس نیست. این طرح به صورت زیر نشان داده می‌شود.

NR	X	$O_۱$
NR		$O_۲$

256 Nonrandom assignment

257 Proxy (توضیح مترجم: متغییر نماینده متغیریست که خود به طور مستقیم مورد نظر محقق نیست اما به جای یک متغیر پنهان یا غیر قابل مشاهده (مورد محاسبه قرار می گیرد. برای آنکه یک متغیر به عنوان نماینده عمل کند، باید رابطه مستقیم (نه الزاماً خطی) با متغیر مورد نظر محقق داشته باشد.)

به عنوان نمونه شارپ و ودربی (Sharp & Wetherbee, 1980) مادرانی که از منافع غذایی پروژه تغذیه زنان، نوزادان و کودکان در می‌سی‌سی‌پی منتفع شده بودند، را با افرادی که از منافع غذایی پروژه برخوردار نشده بودند، مقایسه کردند. نتایج هیچ گونه تفاوت معناداری را از نظر وزن نوزادان در هنگام تولد، یا میزان مرگ و میر نوزادان میان دو گروه مورد مقایسه نشان نداد. با وجود این، احتمال دارد دو گروه موردمقایسه از نظر متغیرهای متعددی از جمله وضعیت غذایی قبلی با یکدیگر متفاوت بوده باشند. این امکان (تفاوت میان گروهها پیش از آزمون) تمیز اثر مداخله از اثرات انتخاب را بسیار دشوار می‌نماید.

از جمله توجیحاتی که برای انتخاب چنین طرح ضعیفی برای انجام مطالعه آورده می‌شود آن است که انجام پیش‌آزمون می‌تواند موجب برانگیختن حساسیت شرکت‌کنندگان شده، و بنابراین نمرات پس‌آزمون آنها را تغییر دهد (ن.ک. اثر آزمون، فصل دوم). اگرچه هنگامی که گروههایی با مداخلههای متفاوت با یکدیگر مقایسه می‌شوند، اندازه‌گیری اثراتی که میان گروهها ثابت هستند، روایی درونی را به خطر نمی‌اندازد. در این مواقع، تنها اثر آزمونهای متفاوت[۲۵۸] می‌تواند خطرساز باشد. این گونه اثرات بعید، اما ممکن هستند. این اثرات را می‌توان به روشهای مختلفی کاهش داد: ۱) در پس‌آزمون، از مقیاس جایگزین مقیاس بکار گرفته‌شده در پیش‌آزمون استفاده کنیم (یک مقیاس در پیش‌آزمون و جایگزین آن در پس‌آزمون)؛ ۲) با استفاده از نظریه پاسخ‌گویه[۲۵۹] تستهای متفاوت را در قالب یک مقیاس واحد همشکل کنیم[۲۶۰]؛ ۳) فاصله زمان میان پیش‌آزمون و پس‌آزمون را طولانی کنیم؛ ۴) از طرح چهارگروهی سولومون برای ارزیابی وجود و تأثیر چنین اثراتی استفاده کنیم؛ ۵) از مقیاسهای نامحسوسی که در مقایسه با روشهای خود-اظهاری کمتر عکس‌العملی هستند، استفاده کنیم؛ ۶) از تکنیکهای مانند کابلهای بُگس[۲۶۱] استفاده کنیم؛ ۷) از پیش‌آزمون پس‌نگر استفاده کنیم ؛ ۸) از گروههای مرجع واضح[۲۶۲] یا معیارهای رفتاری برای تثبیت[۲۶۳] پاسخها استفاده کنیم. بنابراین حتی اگر حساسیت‌زایی آزمونهای متفاوت یک مشکل باشد، هزینه‌های حذف پیش‌آزمون بواسطه لزوم تشخیص اثر سوگیری انتخاب، بیشتر از هزینه مواجهه با آنها از راههای ذکرشده در بالاست.

ارتقاء طرح تنها پس‌آزمون با استفاده از یک پیش‌آزمون مستقل

258 Differential testing effect (توضیح مترجم: فرد بواسطه انجام آزمونهای متعدد قدرت تشخیص بهتری کسب می‌نماید)

259 Item response

260 calibrate

261 Bogus pipeline (توضیح مترجم: در مطالعاتی که با استفاده از ابراز شخصی انجام می‌شود، برای آنکه پاسخگو حقیقت را ابراز کرده و آن را کتمان یا وارونه نکند، کابلهایی به بدن وی نصب می شود. این کابلها در حقیقت کارایی خاصی ندارند اما به پاسخگو گفته می‌شود که این یک دستگاه دروغ سنج است. این باعث می‌شود تا فرد حقیقت را بیان نماید)

262 explicit

263 Anchoring (توضیح مترجم: نوعی سوگیری ذهنی که در آن فرد تکیه زیادی بر اطلاعات اولیه‌ای که دریافت می‌کند می‌نماید)

حتی زمانی که امکان جمع‌آوری اطلاعات و داده پیش‌آزمون (بر روی همان نمونه قبلی و بعد از مداخله) وجود ندارد، همچنان می‌توانیم اطلاعات پیش‌آزمون را از یک نمونه مستقل تصادفی جمع‌آوری نماییم. گروهی که به صورت تصادفی از همان جمعیتی که نمونه پس‌آزمون از آن گرفته شده، انتخاب شده است. اما با این تفاوت که نمونه پیش‌آزمون دارای عضویت بین‌گروهی[۲۶۴] است. این گروهها هنگامی مفید خواهند بود که ۱) مقیاس‌های اندازه‌گیری پیش‌آزمون واکنشی[۲۶۵] است؛ ۲) هنگامی که دنبال کردن همان افراد بسیار گران یا دشوار است؛ و یا ۳) هنگامی که محقق به دنبال مطالعه گروههای دور از دسترس است. طرح مورد بحث را به صورت روبرو نشان می‌دهند:

NR	O_1	X	O_2
NR	O_1		O_2

این طرح مکرراً در بازاریابی، اپیدمولوژی و نظرسنجی‌های سیاسی استفاده می‌شود، و نسبت به طرح پس‌آزمون بدون پیش‌آزمون ارجحیت دارد. این طرح میزان قابل‌توجهی سوگیری ناشی از انتخاب را به تخمینهای مرتبط با اثرات مداخله تحمیل می‌کند، البته اگر گروههای مستقل پیش‌آزمون و پس‌آزمون بطور تصادفی از جمعیت واحدی نمونه‌گیری نشده باشند. سوگیریهای ناشی از انتخاب تنها از طریق نمونه‌گیری تصادفی از بین نمی‌رود. اولاً، هم‌ارز بودن و همسان بودن دو نمونه از طریق نمونه‌گیری تصادفی همچنان محدود به خطای نمونه‌گیری است. بنابراین هنگامی که نمونه‌های کوچک و ناهمگون باشد، قیاس‌پذیری[۲۶۶] بسیار دشوارتر بدست می‌آید. ثانیاً، جمعیتی که نمونه از آن گرفته می‌شود ممکن است در طول زمان بین دو اندازه‌گیری، از نظر ترکیب کیفیتی تغییر کند؛ علی‌الخصوص زمانی که فاصله زمانی میان دو اندازه‌گیری طولانی باشد. این تغییرات کیفی در خصوصیات جمعیت ممکن است به اشتباه به عنوان اثر مداخله قلمداد شود. همچنین، بیشتر تهدیدات متوجه روایی درونی در طرحهای گروه کنترل با نمونه پیش‌آزمون و پس‌آزمون وابسته که در فصل آینده مورد بحث قرار خواهیم داد، در رابطه با نمونه‌های پیش‌آزمون و پس‌آزمون مستقل نیز مصداق دارد. در نهایت، روایی نتایج آماری را می‌توان کاهش داد، زیرا نمونه‌های مستقل در هر موج اندازه‌گیری به عنوان کنترل آماری درون‌گروهی نمونه خود عمل نمی‌کنند. از این رو، این طرحها تنها در زمانی باید بکار گرفته شوند که نیاز به داشتن گروههای مستقل مجاب‌کننده است، و یا مشکلات جدی برای داشتن گروههای وابسته وجود دارد. در صورت لزوم

264 Overlapping membership
265 reactive
266 comparability

بکارگیری این طرح محقق باید توجه جدی به اندازه نمونه، شیوه اجرای طرح نمونه‌گیری، و اینکه چگونه می‌توان قیاس‌پذیری را با استفاده از مقیاسهایی که قابل‌اعتماد و پایدار هستند ارزیابی نمود، داشته باشد.

ارتقاء طرح تنها پس‌آزمون، با استفاده از پیش‌آزمون نماینده[۲۶۷]

یک تکنیک جایگزین، محاسبه نماینده‌هایی برای پیش‌آزمون است- متغیرهایی که به طور مفهومی همبسته و مرتبط با پس‌آزمون درون مداخلهها هستند. این طرح را به صورت زیر نشان می‌دهند، که در آن، A نشان‌دهنده متغیر نماینده، و B نشان‌دهنده پس‌آزمون است.

NR	O_{A1}	X	O_{B2}
NR	O_{A1}		O_{B2}

متغیرهای نماینده باید به صورت مفهومی به متغیر نتیجه مرتبط باشند، نه اینکه تنها متغیری سهل‌الوصول مانند سن، جنسیت، کلاس اجتماعی یا نژاد باشند. برای مثال، هنگامی که کارایی یک واحد درسی ریاضی را برای دانش‌آموزانی که تا به حال ریاضی نخوانده‌اند ارزیابی می‌کنیم، استفاده از نمره ریاضی برای پیش‌آزمون منطقی به نظر نمی‌رسد؛ بنابراین به جای نمره ریاضی بهتر است از نمره استعداد ریاضی یا استعداد جبر استفاده نماییم. هر چه میزان همبستگی میان پس‌آزمون و متغیر واسطه پیش‌آزمون بیشتر باشد، احتمال این که بتوانیم با استفاده از نتایج پیش‌آزمون، میزان سوگیری در انتخاب نمونه، و اینکه گروهها قبل از آزمون و مداخله با یکدیگر تفاوت معناداری داشته‌اند را شاخص‌گذاری کنیم، بیشتر می‌شود. این همبستگی همچنین می‌تواند سوگیری ناشی از ریزش را نیز که نشان‌دهنده آن است که تا چه میزان کسانی که مطالعه را ترک کرده‌اند، از کسانی که در مطالعه مانده‌اند متفاوتند، را مشخص نماید[۲۶۸]. با وجود این، به کارگیری این روش نتایج ضعیفتری نسبت به حالتی بدست خواهد داد که خود متغیر نتیجه (که در پس‌آزمون نیز اندازه‌گیری می‌شود) در پیش‌آزمون مورد اندازه‌گیری قرار بگیرد.

ارتقاء طرح تنها پس‌آزمون با استفاده از جفت کردن[۲۶۹] و طبقه‌بندی کردن[۲۷۰]

نبود پیش‌آزمون موجب عدم وجود اطلاعات نسبت به وجود سوگیری در انتخاب نمونه می‌شود. محققین معمولاً سعی می‌کنند این مشکل را بوسیله شکل‌دادن به گروههای کنترل، و درمان از طریق جفت کردن یا طبقه‌بندی بر اساس متغیرهای همبسته با متغیر پس‌آزمون، کاهش دهند.

267 proxy
268 Index
269 Matching
270 Stratifying

تعاریف. هنگام جفت‌سازی، محقق افراد با نمرات مشابه از نظر متغیر جفت‌سازی را به گروههای کنترل و آزمون تخصیص می‌دهد. بنابراین هر دو گروه کنترل و مداخله افرادی را شامل می‌شوند که از نظر متغیر جفت‌شده با یکدیگر شبیه هستند. مثلاً، در مطالعه‌ای اثر برنامه اطلاعات تغذیه بر فروش محصول و سهم بازار در سوپر مارکتهای تحت‌مالکیت یک زنجیره با شیوه‌های مدیریتی واحد مورد مطالعه قرار گرفت. برای این مطالعه ده فروشگاه در واشنگتن و ده فروشگاه در مریلند انتخاب شدند. این فروشگاهها از نظر مشخصات اندازه فروشگاه، مشخصات اجتماعی-اقتصادی مشتریان جفت شدند، چون این متغیرها روی میزان فروش و اندازه بازار تأثیرگذار بود.

مطالعات بر روی دوقلوها نمونه بسیار ویژه‌ای برای جفت‌سازی است. فرض بر این است که میزان شباهت دوقلوها به یکدیگر بیشتر از شباهت میان آنها و دیگران است. این موضوع برای دوقلوهای همسان بیشتر صدق می‌کند، زیرا ساختار ژنتیکی یکسانی دارند. اما دوقلوهای غیرهمسان تنها در بخشی از ساختار ژنتیکی مشترک هستند. به علاوه، دوقلوها عموماً در معرض اثرات محیطی مشابهی قرار می‌گیرند. برای مثال، در یک خانواده بزرگ می‌شوند، توسط یک پدر و مادر مشترک با زمینه اقتصادی و اجتماعی یکسان. تمامی این شباهتهای ژنتیکی و محیطی باعث می‌شود استفاده از دوقلوها به عنوان جفتها در مطالعات شبه‌آزمایشی رویکردی بسیار قوی به حساب بیاید.

از دیگر تکنیکهای بسیار نزدیک به جفت‌سازی می‌توان به طبقه‌بندی[۲۷۱] اشاره کرد. این تکنیک هنگامی به کار گرفته می‌شود که افراد در مجموعه‌های همگنی قرار دارند که دارای تعداد افرادی بیشتر از تعداد شرایط آزمایش است. یک مثال در این خصوص، طبقه‌بندی بر مبنای جنسیت است. غیرممکن است بتوان برای یک مرد، جفت یا همسازی بهتر از مردی دیگر (در میان گروه بزرگی از مردان) یافت. بنابراین یک بلوک دارای تعداد خیلی بیشتری مرد، در مقایسه با تعداد شرایط آزمایش، خواهد بود. برخی اوقات یک طبقه با استفاده از متغیرهای پیوسته ساخته می‌شود. برای مثال، با استفاده از میانه نمرات آزمون رانندگی، گروه را به دو دسته بزرگ تقسیم می‌کنیم. در نتیجه این طبقه‌بندی، گروههای ایجاد شده نسبت به حالتی که شرکت‌کنندگان جفت می‌شدند، کمتر همگن و یکدست خواهد بود. اگر لازم است از طبقه‌بندی استفاده شود، تعداد بیشتر طبقات عموماً از تعداد کمتر بهتر است؛ و پنج طبقه غالباً برای حذف ۹۰٪ واریانسی که به واسطه جفت‌سازی کاهش بیابد کافی خواهد بود (Cochran, 1986). در ادامه به بحث درباره جفت‌سازی و روشهای آن بحث خواهیم کرد. اما نکاتی که ذکر خواهد شد در مورد طبقه‌بندی نیز کاربرد دارد. بعلاوه برخی از روشهایی که مورد بحث قرار خواهد گرفت (مانند جفت‌سازی بهینه) تمایز مفهومی میان جفت‌سازی و طبقه‌بندی را کم رنگتر می‌کند، اگرچه کاربردهای عملی باید همچنان واضح باشد.

271 Stratifying

روشهای جفت‌کردن. روشهای متنوعی برای جفت‌کردن مورد استفاده قرار می‌گیرد (Cochran, 1983; Cochran & Rubin, 1973; Costanza, 1995; Dehejia & Wahba, 1999; Gu & Rosenbaum, 1993; Heckman, Ichimura, & Todd, 1997; Henry & McMillan, 1993; Marsh, 1998; Rosenbaum, 1995a; H.Smith, 1997). ۱) جفت‌کردن دقیق[۲۷۲]: لازم است که افراد در دو گروه جفت‌شده دارای نمرات دقیقاً مشابه باشند. اگرچه اگر نمونه کوچک باشد، احتمال پیدا کردن افرادی با نمرات دقیقاً مشابه کمتر می‌شود. همینطور اگر توزیع شرکت‌کنندگان در گروهها نامتوازن باشد، و یا اینکه مقیاس ارزیابی و اندازه‌گیری در مورد متغیر جفت‌سازی (متغیری که بر مبنای آن افراد و گروهها جفت می‌شوند) نمره‌دهی خیلی نزدیک به هم[۲۷۳] استفاده کرده باشد، ۲) جفت‌کردن کولیس[۲۷۴]: نمرات در این نوع جفت‌سازی لازم نیست کاملاً مشابه باشند، بلکه باید فاصله یکسان نسبت به یکدیگر قرار داشته باشند. روشهای متفاوتی برای جفت کردن این فاصله وجود دارد، مانند روش نزدیکترین همسایه یا جفت‌کردن بر مبنای فاصله ماهالانوبیس (Hill, Rubin, & Thomas, 2000; Rosenbaum, 1995a). ۳) برخی اوقات تعداد افراد موجود در گروه کنترل بیشتر از گروه آزمون است، بنابراین اگر محقق بتواند کنترلهای متعددی انتخاب نماید (Henry & McMillan, 1993)، ممکن است بتوان جفت‌سازی و توان آماری را ارتقاء داد. برای مثال، جفت‌سازی شاخص[۲۷۵]، تعداد متعددی از افراد گروه کنترل را در دو طرف (بالاتر و پایین‌تر) افراد گروه آزمون انتخاب می‌کند. در جفت‌سازی گروه خوشه‌ای[۲۷۶]، از تحلیل خوشه‌ای استفاده می‌شود تا گروه آزمون را در خوشه‌ای از افراد کنترل مشابه جای دهد. ۴)جفت‌سازی گروه معیار[۲۷۷]: افرادی از گروه کنترل را انتخاب می‌کند که بر اساس مقیاس فاصله چندمتغیره، در نزدیکی افراد گروه آزمون قرار داشته باشند؛ در نهایت، ۵) در جفت‌سازی بهینه[۲۷۸] هر فرد در گروه آزمون ممکن است جفتهای متعددی در گروه کنترل داشته باشد (و یا بالعکس). همچنین لازم است تا نقاط قوت و ضعف هر یک از روشهای جفت‌سازی و همچنین قیاس این روش‌ها با یکدیگر و دیگر گزینه‌ها مانند تحلیل کوواریانس به صورت تفصیلی انجام شود[۲۷۹].

مشکلات ناشی از جفت‌سازی. جفت‌سازی سرگذشت پرفراز و نشیبی در شبه‌آزمایشها دارد، چون همواره امکان بروز سوگیری انتخاب بواسطه انجام جفت‌سازی وجود دارد. کمترین ریسک در این زمینه، انجام جفت‌سازی به طور ناقص[۲۸۰] است؛ زیرا برخی از متغیرهای غیراضافی که نقش مهمی در تعیین نتایج دارند، در روش

272 Exact matching
273 Fine
274 Caliper
275 Index
276 Cluster group matching
277 Benchmark
278 Optimal
۲۷۹ برخی از مزایای جفت‌سازی و آنالیز کوواریانس در پیوست فصل بعدی و در فصل مربوط به طرحهای آزمایشی مورد بحث قرار خواهد گرفت.
280 undermatching

جفت‌سازی لحاظ نمی‌شوند. برای مثال، لوی و همکارانش (Levy et al., 1985) فروشگاهها را برمبنای دو متغیر جفت کردند، در حالی که دیگر متغیرها مانند خط محصول فروشگاه، و نزدیکی به دیگر فروشگاهها می‌توانست به نحو بهتری میان فروشگاهها تمایز ایجاد کند، و با نتایج همبستگی داشته باشد. اگر متغیرهای جفت‌سازی اضافی بکار گرفته می‌شد، سطح بالاتری از همترازی و شباهت میان فرشگاههای آزمون و کنترل بدست می‌آمد. علاوه بر این، از آنجا که جفت‌سازی هیچگاه نمی‌تواند از نظر متغیرهایی که در جفت‌سازی به کار گرفته نشده‌اند هم‌ارزی ایجاد کند، سوگیری انتخاب را هیچگاه نمی‌توان با اطمینان از میان برد.

اگرچه علّت نگاه بدبینانه موجود نسبت به جفت‌سازی کمتر به دلیل احتمال انجام ناقص جفت‌سازی است، زیرا هرچه باشد فرد را یک قدم به پاسخ صحیح نزدیکتر می‌سازد. بلکه علّت عمده این بدبینی آن است که جفت‌سازی می‌تواند نتایجی تولید کند که در مقایسه با حالتی که جفت‌سازی انجام نمی‌گرفت، از پاسخ صحیح، فاصله بیشتری داشته باشد (یعنی جفت‌سازی ما را از پاسخ صحیح دور می‌سازد). مطالعه کمپبل و ارلباچه (Campbell & Erlebacher, 1970) نشان داد که چطور جفت‌سازی می‌تواند چنین نتایجی را به بار بیاورد. این دو محقق با هدف بررسی علل بدست آمدن نتایج غیرمنتظره از ارزیابی یک برنامه آموزش پیش‌دبستانی دست به انجام مطالعه‌ای زدند. نتایج بدست آمده نشان می‌داد که کودکان انتخاب‌شده برای این برنامه (بر خلاف انتظار) نسبت به گروه جفت‌شده عملکرد ضعیف‌تری داشتند. بررسی نشان داد که جمعیتهایی که جفتها از آنها گرفته شده بودند از نظر متغیر جفت‌سازی به طور کامل همپوشانی نداشتند، و بنابراین جفتهای انتخاب‌شده برای برنامه پیش‌دبستانی در مقایسه با گروه کنترل خود، از سر متفاوتی از طیف توزیع خود انتخاب شده بودند (گروه آزمون از سر بالایی طیف و گروه کنترل از سر پایینی طیف توزیع خود انتخاب شده بودند). اگر متغیر جفت‌سازی با خطا محاسبه شود یا به طور ناکاملی با نتایج همبستگی داشته باشد، [سوگیری] رگرسیون آماری رخ خواهد داد، که نتایج برنامه پیش‌دبستانی را حتی مضرّ و آسیب‌زا نشان می‌دهد.

مارش (Marsh, 1998) مثال مشابهی را در مورد ارزیابی برنامه‌های طراحی‌شده برای کودکان بااستعداد و تیزهوش ارائه می‌کند، که در آن کودکان نسبت به گروه کنترل از عملکرد بهتری برخوردار بودند. ممکن است جفت‌شده‌های گروه کنترل از محوطه همپوشانی میان انتهای پایینی توزیع جمعیتی گروه آزمون (انتهایی که در آن خطاهای منفی بیشتری وجود دارد)، و انتهای بالایی توزیع جمعیتی گروه کنترل (انتهایی که در آن خطاهای مثبت بیشتری وجود دارد)، گرفته شده باشند. در پس‌آزمون، کودکان گروه آزمون نسبت به میانگین جمعیت خود بهبود در عملکرد داشتند، در حالی گروه کنترل نسبت به میانگین جمعیت خود تنزل کرده بودند. سوگیریهای حاصل از فرایند انتخاب جفتها باعث می‌شود تا برنامه‌های کودکان تیزهوش به نظر کارآمد بیاید حتی اگر در واقع این طور نبوده باشد.

اصول راهنما برای انجام بهتر جفت سازی. درس مهمی که می‌توان از مثالهای آورده‌شده در سطور قبلی گرفت، این است که اگر جفت‌سازی با استفاده ازمتغیرهای ناپایدا و غیرقابل اعتماد انجام شود، و یا زمانی که گروههای غیرهم‌ارزی که جفتها از آنها گرفته می‌شوند در زمان جفت‌سازی به طور فزاینده‌ای غیرمشابه باشند، جفت‌سازی در شبه‌آزمایشها از کمترین کارایی برخوردار خواهد بود (و می‌تواند به جای مفیدبودن آسیب‌زا باشد). اولین اصل آن است که گروههایی را انتخاب کنیم که تا حد امکان قبل از انجام جفت‌سازی به یکدیگر شبیه باشند (البته تا آنجا که سؤال و زمینه تحقیق اجازه می‌دهد). اگر توزیعهای دو گروه در متغیر جفت‌سازی به میزان زیادی همپوشانی داشته باشند، می‌توان بدون آنکه نیاز به گرفتن جفت از سرهای مخالف و انتهای توزیعها وجود داشته باشد، جفتهای زیادی انتخاب کرد. برای مثال، اگر گروه کنترل از میان افرادی انتخاب شود که واجد شرایط بودن در گروه آزمون بوده‌اند، اما مثلاً دیر ثبت‌نام کرده‌اند (در مقایسه با حالتی که افراد گروه کنترل واجد شرایط بودن در گروه آزمون نبوده باشند)، گروههای غیرهم‌ارز احتمالاً همپوشانی بیشتری خواهند داشت. هنگامیکه امکان چنین انتخابی وجود نداشته نباشد، بررسی همپوشانی توزیع دو گروه می‌تواند محققین را نسبت به احتمال بروز [سوگیری] رگرسیون میان جفتها هوشیار نماید.

روش دوم، استفاده از متغیرهایی برای جفت‌سازی است که پایدار و دارای پایایی باشند. برخی متغیرها مانند جنسیت و سن با میزان اندکی خطا محاسبه می‌شوند، و بنابراین متغیرهای خوبی برای جفت‌سازی هستند، البته *به شرطی* که با نتایج همبستگی داشته باشند. پایایی دیگر متغیرهای جفت‌سازی را غالباً می‌توان از طریق تجمیع[۲۸۱] ارتقاء داد؛ برای مثال، از طریق ترکیبی از متغیرهای پیش‌آزمون متعددی که بطور همزمان بکار گرفته می‌شوند (رویکرد نمرات رغبت[۲۸۲] که در فصل پنج توضیح داده خواهد شد، نمونه‌ای از این روشهاست)، و یا از طریق ساختن ترکیبی از افراد (استفاده از میانگین مدارس بجای داده‌های انفرادی دانش‌آموزان)، و در نهایت، از طریق میانگین گرفتن از دو یا چند پیش‌آزمون بجای استفاده از تنها یک پیش‌آزمون. با توجه به اینکه میانگین گرفتن از مشاهدات بیشتر خطاهای تصادفی را خنثی می‌کند، روش آخر می‌تواند مانع از انتخاب افراد به این دلیل که نمرات پیش‌آزمون آنها بسیار بزرگ یا بسیارکوچک بوده شود (مسأله‌ای می‌تواند به بروز سوگیری رگرسیون آماری منتهی شود). برای مثال، شیوه جفت‌سازی بکار گرفته‌شده در مطالعه اخیر میلسپ و همکارانش (Millsap, et al., 1997) که درباره اثر برنامه ارتقاء در مدارس بر دستاوردهای دانش‌آموزان در دیترویت صورت گرفت را می‌توان با دیدی خوشبینانه نگریست. در این مطالعه، از روش محاصره جفتهای پایدار که در آن ۱۲ مدرسه [در شرایط] آزمون، با ۲۴ مدرسه جفت‌شده مقایسه شدند، استفاده شد. جفت‌سازی بر مبنای سه روش انجام شد: (۱) بر مبنای متغیرهایی که به شیوه‌ای پایا و قابل‌اعتماد محاسبه شده بودند (در این نمونه خاص

281 aggregation
282 Propensity score

موقعیت مکانی ناحیه فرعی یا زیربخش، نمرات آزمون دستاوردها در سطح مدارس، و ترکیب نژادی)؛ (۲) از طریق میانگین‌گیری نمرات چندین سال به جای استفاده از نمرات یک سال؛ و (۳) با استفاده از تجمیع (مثلاً استفاده از نمرات مدارس) به جای استفاده از نمرات انفرادی. بعلاوه، از میان مجموعه‌ای چهار تا شش جفت برای هر مدرسه آزمون، دو مدرسه برای مقایسه انتخاب شدند، تا مدرسه آزمون را محاصره نمایند: یکی از این مدارس قدری بهتر از مدرسه آزمون عمل می‌کرد، و دیگر سطح عملکردی پایینتر از آن داشت. استفاده از دو مدرسه برای مقایسه با هر مدرسه آزمون توان آماری را با هزینه کمتری (نسبت به افزایش توان از طریق افزایش تعداد مدارس آزمون) افزایش می‌داد. زیرا نیازی به اجرای مداخله پرهزینه در مدارس کنترل نداشت. این افزایش توان آماری، بویژه در زمانهایی که نمرات تجمعی (نمرات در سطح مدارس برای مثال) مورد مطالعه قرار می‌گیرند و همچنین زمانی که تعداد معدودی تجمیع در اختیار داریم، از اهمیت خاصی برخوردار است.

در مواقع جفت‌سازی برمبنای بیش از یک متغیر به طور همزمان، هرچه تعداد متغیرها بیشتر باشد، جفت‌سازی دشوارتر خواهد بود. با این حال کاهش این متغیرها به یک ترکیب چند-متغیری جفت‌سازی را مقرون به صرفه‌تر و شدنی‌تر می‌سازد (Henry & MaMillan, 1993). جفت‌سازی فاصله‌ای چندمتغیری که پیش از این توضیح داده شد، از چنین ترکیبی استفاده می‌کند، همانطور که جفت‌سازی برمبنای نمرات رغبت نیز این کار را انجام می‌دهد. نمره رغبت به منظور پیش‌بینی عضویت در گروه، با استفاده از رگرسیون لجستیک، و بر مبنای متغیرهای تعیین‌کننده عضویت بدست می‌آید. جفت‌سازی بر مبنای نمره رغبت تفاوتهای میان‌گروهی را در طول تمامی متغیرهای مشاهده‌شده بکار رفته در معادله نمره رغبت، به حداقل می‌رساند. این جفت‌سازی همچنین برای انواع « شکلهای کارکردی[۲۸۳]» مابین نمره رغبت و نتایج، مستحکم است. جفت‌سازی بر مبنای نمرات رغبت را می‌توان با استفاده از تحلیلهای حساسیتی که به منظور بررسی میزان استحکام اثر مشاهده‌شده در برابر سوگیریهایی با اندازه مشخص طراحی شده‌اند، تکمیل نمود. نمرات رغبت و تحلیل سوگیریهای پنهان به تفصیل در پیوست فصل آینده مورد بحث قرار خواهد گرفت.

برای انجام جفت‌سازی در شبه‌آزمایشها، باید بر مشکلات جدی متعددی غلبه کنیم. جفت‌سازی می‌تواند تنها بر مبنای متغیرهای قابل‌مشاهده انجام گیرد، و بنابراین، سوگیریهای پنهان همچنان باقی بمانند. کاهش عدم‌پایایی احتمال تهدید رگرسیون یا بازگشت به میانگین را کاهش می‌دهد. اما اگر متغیرهای جفت‌سازی و متغیرهای نتایج - علیرغم محاسبه کامل و دقیق - به خوبی همبستگی نداشته باشند، این تهدید همچنان می‌تواند رخ دهد (Campbell & Kenny, 1999). برخی تهدیدات نسبت به روایی درونی نیز بعد از پیش‌آزمون اتفاق می‌افتند (مانند گذشت زمان). هیچ جفت مستقیمی برای این در برنامه‌ریزی طرح آزمایش وجود ندارد. اگرچه ما همچنان بر این باور هستیم که رویه‌های جفت‌سازی بهتری که در این بخش ارایه شد، از رویکردهای ساده‌انگارانه به مقوله

283 Functional form

جفت‌سازی، که بر مبنای یک متغیر و یا بر پایه جمعیتهای متفاوت عمل می‌کنند، موفق‌تر هستند. محققینی که از جفت‌سازی بهره می‌گیرند، می‌بایست از منافع این رویه‌های جامع‌تر بهره ببرند. دوران جفت‌سازیهای ساده با یک متغیر، که آن هم به طور غیرقابل اعتمادی محاسبه می‌شد، باید به کلی پشت‌سر گذاشته شود.

ارتقاء طرح تنها پس‌آزمون، با استفاده از کنترلهای درونی

گروههای کنترل درونی به طور منطقی از جمعیتی شبیه به جمعیتی که گروه آزمایش از آن گرفته شده، انتخاب می‌شوند. برای مثال ایکن و همکارانش (Aiken et al., 1998) از یک طرح تصادفی و یک شبه‌آزمایش به طور همزمان برای آزمون اثرات یک برنامه اصلاح نوشتار بهره بردند. گروه کنترل درونی غیرهم‌ارز آنها متشکل از دانشجویان واجد شرایطی بود که خیلی دیر برای آزمایش ثبت‌نام کرده بودند (گروهی که به طور قابل‌توجهی مشابه به دانش‌آموزان واجد شرایطی بودند که به موقع ثبت نام کرده بودند). در چنین مواردی احتمال سوگیری ناشی از انتخاب در مقایسه با مواردی که گروههای کنترل خارجی باشند (مثل انتخاب دانش‌آموزان از یک دانشگاه دیگر، یا کسانی که نمرات ACT آنها پایینتر از حد مجاز بوده و واجدشرایط برنامه نیستند) کمتر است. کنترلهای درونی الزاماً تضمین‌کننده شباهت نیستند. برای مثال، در برخی تحقیقات رواندرمانی که از پذیرندگان درمان به عنوان گروه آزمون، و از ردکنندگان درمان یا ریزشها، به عنوان گروه کنترل استفاده می‌کنند. در این حالت، احتمال بروز مشکلات بدیهی ناشی از انتخاب وجود دارد. مشکلاتی که می‌توانند تخمینهای مرتبط با اثر درمان را تحت تأثیر قرار دهند. بکر و رودریگزر (Becker & Rodriguez, 1979) یک گروه کنترل درونی را برای مطالعه اثر ارجاع متهمان دادگاههای جنایی از سیستم حقوقی به خدمات آموزشی و اجتماعی، مورد استفاده قرار دادند. وکلای حقوقی نسبت به تخصیص تصادفی در زمینه‌های حقوقی اعتراض داشتند. اگرچه، تعداد افراد ارجاع داده‌شده دو برابر تعداد قابل‌قبول بود. بنابراین بکر و رودریگرز از یک طرح تخصیص دو مرحله‌ای استفاده کردند که در آن، (۱) زمان، به بلوکهایی ۱۱، ۱۳، ۱۵، ۱۷، ۱۹ و ۲۱ ساعتی (از نظر مدت‌زمان) تقسیم شده بود، و (۲) ۵۰٪ اول مشتریان موردانتظار در طول هر بلوک به گروه آزمایش تخصیص داده می‌شدند، و باقی‌مانده‌ها به گروه کنترل تخصیص می‌یافتند. کارکنان برنامه از تعداد ساعتهای بلوک در حال اجرا بی‌اطلاع بودند. بنابراین آنها نمی‌توانستند به سادگی پیش‌بینی کنند که چه زمانی سهم ۵۰٪ پر می‌شود، و یا اینکه مراجع بعدی به گروه آزمایش تخصیص می‌یابد یا خیر. همچنین دادگاهها نمی‌توانستند به طور ترجیحی برخی از مراجعین را به گروه آزمایش هدایت کنند، چون نمی‌دانستند چه بلوکی در حال انجام است، و همچنین افراد از دادگاههای متعدد ارجاع داده می‌شدند. بررسی تفاوت میان گروهها نشان داد که شرایط آزمون و کنترل از نظر متغیرهای محاسبه‌شده در مرحله پیش‌آزمون، مشابه بودند. احتمالاً مشخصه زمان تصادفی می‌تواند احتمالی بودن تخصیص را افزایش دهد، البته اگر کارمندان از زمان شروع بلوک زمانی جدید بی‌اطلاع باشند، و اگر کارکنانی غیر از کارکنان برنامه فرایند تخصیص را به عهده بگیرند.

اگرچه، برخی مواقع ممکن بود این شرایط نقض شود، به این معنی که برخی کارکنان مراجعان را به سمت شرایط آزمون هدایت می‌کنند. اما در نهایت، بزرگی هر قسمت همواره بیشتر از یک بود. بنابراین کارکنان می‌دانستند که هنگامی که یک مراجع به گروه آزمون ارجاع داده می‌شود، بعدی نیز به همین گروه ارجاع داده می‌شود. هنگامی که کارکنان برنامه در مورد تخصیص بعدی مطلع بودند، تخصیص می‌توانست به سادگی دچار سوگیری شود (Chalmers, Celano, Sacks, & Smith, 1983; Dunford, 1990). این رویه همچنان نیازمند آن بود که تعداد ارجاعها بیشتر از ظرفیت برنامه باشد. این شرط موقعیتهای مناسب برای بکارگیری این طرح را محدود می‌کند. اما همچنان این رویه برای مواقعی که فرصتها و گزینه‌های دیگر انجام‌پذیر نیستند، قابل‌بکارگیری است. این رویه به دو طرح بسیار قوی که در آینده به آنها خواهیم پرداخت شباهت دارد: (۱) طرح ناپیوستگی رگرسیون[۲۸۴] که در آن ترتیب ارجاعها، متغیر انتخاب است، و (۲) مطالعه‌ای با تخصیص تصادفی بلوکهای زمانی به شرایط.

ارتقاءطرح تنها پس‌آزمون، با استفاده از گروههای کنترل متعدد

همانطور که در بحث مرتبط با جفت‌سازی عنوان شد، اغلب مواقع امکان استفاده از گروههای کنترل غیرهم‌ارز وجود دارد. برای مثال بل و همکارانش (Bell et al., 1995) یک گروه آزمون آموزش ضمن‌خدمت را با چهار گروه کنترل مقایسه کردند، که عبارت بودند از، آنها که نتوانسته بودند برای برنامه ثبت‌نام کنند، آنها که درخواست آنها رد شده بود، درخواستهای پذیرفته شده‌ای که نتوانسته بودند آموزش (مداخله) را آغاز کنند، و در نهایت، آنها که مداخله (آموزش) را آغاز کرده اما نتوانسته‌اند به پایان ببرند. استفاده از گروههای کنترل متعدد، از چند جهت مفید است. اگر گروههای کنترل به همان میزان که از گروه آزمون تفاوت دارند، با یکدیگر نیز متفاوت باشند، بدیهی است که تفاوتهای مشاهده‌شده در نتایج، نمی‌تواند به دلیل مداخله باشد. بنابراین، تفاوت در نتایج می‌تواند نشان‌دهنده قدرت و اهمیت سوگیریهای پنهانی موجود باشد. اگر جهت سوگیری در هر گروه کنترل از پیش معلوم است، می‌توان اثر مداخله را از میان سوگیریهای از پیش شناخته‌شده براکت[۲۸۵] کرد. برای مثال، کمپبل (Campbell 1969a) در مورد کنترلهای واریاسیونهای سیستماتیک و کنترل از طریق براکت کردن مباحثی را ارائه می‌کند. در مورد کنترل واریاسیونهای سیستماتیک، محقق خطر اصلی برای روایی را تشخیص می‌دهد، و سپس کنترلهای متعددی را که دامن قابل‌قبول آن تهدید را محاسبه می‌کنند، انتخاب می‌کند. اگر اثرات مشاهده‌شده در طول آن طیف نوسان نداشته باشند، موجّه‌بودن و امکانپذیزی آن تهدید کاهش می‌یابد. در روش براکت کردن، محقق دو گروه انتخاب می‌کند، یکی گروهی که انتظار می‌رود در صورت بی‌اثر بودن مداخله

284 Regression discontinuity
285 Bracket

از گروه آزمون عملکرد بهتری داشته باشد، و دیگری گروهی که عملکردی ضعیفتر از گروه آزمون دارد. اگر گروه آزمون بهتر از هردو این گروههای کنترل عمل نماید، استنباطهای علّی تقویت می‌شود.

روزنباوم مثال مشابهی را در مورد استفاده از کنترلهای متعدد ذکر می‌کند. زابین و همکارانش (Zabin et al., 1989) این کنترلها را در مورد بررسی اثرات مترتب بر زنان نوجوان سیاه پوستی که سقط جنین داشته‌اند بکار می‌گیرند. زابین و همکارانش زنان حامله‌ای که سقط جنین کرده بودند را با زنانی که فرزند داشتند، و زنانی که به دنبال سقط بودند اما پس از آزمایش متوجه شدند که اصلاً حامله نبوده‌اند، مقایسه کردند. یافته‌های آنها نشان داد که زنان گروه آزمون (زنانی که سقط داشته‌اند) نسبت به دو گروه دیگر، پس از دو سال، نتایج تحصیلی بهتری کسب کردند. این موضوع نشان‌دهنده نتایج تحصیلی بهتر در میان گروههایی بود که از نظر آموزشی مشابه بودند. اگر تنها یک گروه کنترل وجود می‌داشت، و آن گروه زنان دارای فرزند می‌بود، ممکن بود اینطور نتیجه‌گیری شود که نتایج بهتر تحصیلی زنان با سقط جنین، به دلیل مسئولیتهای نگهداری فرزند در زنان دارای فرزند بوده است. اما این نتیجه‌گیری را نمی‌توان در مورد گروه کنترل زنانی که از ابتدا حامله نبوده‌اند، و در نتیجه مسئولیت نگهداری از کودک را بر عهده نداشته‌اند، مطرح نمود. نتایج این آزمایش طرح این بحث که سقط جنین باعث کاهش دستاوردهای تحصیلی می‌شود را بلاموضوع می‌ساخت.

ارتقاء طرح تنها پس‌آزمون، با استفاده از یک اثر تعامل پیش‌بینی‌شده[۲۸۶]

برخی اوقات نظریه مبنای[۲۸۷] طراحی فرضیات آنچنان از قوام و پیچیدگی لازم برخوردار است که می‌تواند فرضیات علّی کاملاً متمایزی تولید کند؛ فرضیاتی که در صورت تأیید، می‌توانند نقض‌کننده بسیاری از تهدیدات روایی درونی باشند. زیرا متغیرهای تهدیدکننده روایی درونی قادر به ایجاد چنین اثرات پیچیده‌ای نیستند. یک مثال برای این موضوع، شبه‌آزمایش انجام شده توسط سیور (Seaver, 1973) در مورد اثرات انتظارات عملکردی معلّم بر موفقیت تحصیلی دانش‌آموزان است. سیور دانش‌آموزانی که خواهران یا برادران آنها نمرات و موفقیتهای بیشتر (کمتری) کسب کرده بودند را انتخاب کرد. او این دو گروه (بیشتر در مقابل کمتر) را به دو دسته تقسیم کرد: یک از این دسته‌ها همان معلمی را داشتند که به خواهران و برادران بزرگتر آنها درس داده بود، و دسته دوم معلم متفاوتی داشتند. سیور پیش‌بینی کرد که انتظارات معلم باعث می‌شود کودکانی که خواهران و برادرانشان عملکرد بهتری داشته‌اند، نسبت به دیگر کودکان عملکرد بهتری داشته باشند. این تفاوت عملکردی در گروهی که معلمی یکسان با خواهران و برادرانشان داشته‌اند، بیشتر از گروهی است که معلم متفاوت داشته‌اند. به بیان دیگر، نتایج، فرضیات مرتبط با اثر برهم‌کنش را تأیید می‌کرد. در این مورد خاص، اثر برهم‌کنشی پیش‌بینی شده مفید بود، چون (۱) نظریه مبنا یک الگوی پیچیده از داده‌ها را پیش‌بینی می‌کرد؛

286 predicted interaction

287 Substantive theory

(۲) گروههای کنترل خواهران و برادران در دسترس بودند. اگرچه غیرهم‌ارز بودند، اما به طور قابل‌قبولی از نظر بسیاری از سوابق و مشخصات خانوادگی مشابه بودند؛ (۳) محاسبه متغیر وابسته (موفقیت تحصیلی) به صورت قابل‌اتکایی محاسبه شد، و (۴) اندازه بزرگ نمونه امکان انجام آزمون پرقدرت برای اثر برهم‌کنش را فراهم می‌ساخت.

اگرچه محقق باید حتی زمانی که اثر برهم‌کنش تأیید می‌شود، همچنان محتاط باشد. ریچارت (Reichardt, 1985) نشان داد که نتایج سیور ممکن است نوعی نتایج مصنوعی رگرسیونی[۲۸۸] باشد. سیور همزمان با تقسیم دانش‌آموزان به چهار گروه، معلمان را نیز دسته‌بندی کرد. فرض کنید خواهران و برادران بزرگتر با عملکرد بالاتر از متوسط، معلمی داشته‌اند که تواناییها (و نه انتظارات) بالاتری داشته است. خواهر و برادر جوانتری که به همان معلم سپرده می‌شود، نیز قاعدتاً تعلیمات بهتری دریافت می‌کند. اما خواهر و برادر جوانتری که به معلم دیگری سپرده می‌شود، ممکن است تعطیلات در سطح متوسطی دریافت کند. همین مسأله را می‌توان در مورد خواهران و برادران بزرگتری که عملکرد ضعیف‌تر داشته‌اند نیز گفت. تفاوتها در اثربخشی معلمها می‌تواند به حساب اثر سرریزشده[۲۸۹] که سیور پیدا کرده است، گذاشته شود. اگرچه دیگر تهدیدات مرتبط با روایی درونی به نظر بی‌اثر شده، و اگر نمرات ارزیابی توانایی معلمین در دسترس باشد، این تهدید را نیز می‌توان به طور کاربردی ارزیابی نمود.

ارتقاء طرحهای بدون گروه کنترل، با استفاده از ساختن متضادهایی[۲۹۰] غیر از موارد گروههای کنترل مستقل

هنگامی که جمع‌آوری داده‌ها در مورد انواع کنترلهای مستقل که در قسمت قبلی مورد بحث قرار گرفت وجود ندارد، برخی مواقع می‌توانیم متضادهایی بسازیم که شبیه عملکرد یک گروه کنترل مستقل رفتار می‌کنند. سه نمونه از این متضادها عبارتند از (۱) چندبعدی‌سازی رگرسیونی[۲۹۱]، که مقادیر پس‌آزمون واقعی و انعکاس‌یافته را مقایسه می‌کند؛ (۲) مقایسه نرمال‌شده، که دریافت‌کنندگان مداخله را با نمونه‌های نرمال مقایسه می‌کند؛ (۳) داده دست‌دوم، که دریافت‌کنندگان مداخله را با نمونه‌های گرفته‌شده از داده‌های جمع‌آوری شده قبلی (مانند پیمایشهای در سطح ملّی) مقایسه می‌کند. تمامی این گزینه‌ها نقاط ضعف متعددی دارند. بنابراین ترجیح بر این است که این تضادهای مصنوعی اصلاً به کار گرفته نشوند. اگرچه این راه‌حلها غالباً راحت و ارزان هستند، بنابراین یک مطالعه می‌تواند با هزینه‌ای اندک این گزینه‌ها را با دیگر عناصر طرح شبه‌آزمایش ترکیب نموده، و دیگر گزینه‌های تبیین اثرات را بررسی نماید.

288 artifact
289 Crossover
290 Contrasts
291 Extrapolation

تضادهای چندبعدی‌سازی رگرسیونی

این طرح نمرات به دست‌آمده از گروه آزمون در مرحله پس‌آزمون را با نمراتی که پیش‌بینی می‌شد براساس دیگر اطلاعات بدست بیاید، مقایسه می‌کند. برای مثال، کوک و همکارانش (Cook et al., 1975) اثرات تماشای برنامه تلویزیونی «خیابان کنجد» را در نقاط مختلف کشور موردمطالعه قرار دادند. ابتدا یک پیش‌آزمون انجام شد، شش ماه بعد یک پس‌آزمون. سن (به ماه) در مرحله پیش‌آزمون به منظور پیش‌بینی موفقیت تحصیلی در مرحله پیش‌آزمون مورد استفاده قرار گرفت. نتایج این محاسبه تخمینی از میزان موفقیت موردانتظار به ازای هر ماه سن کودک به دست می‌داد. تخمین تغییرات ماهانه، سپس به عنوان ابزار پیش‌بینی میزان دستاورد هر کودک در اثر بلوغ در طی شش ماه بکار گرفته شد، و پیش‌آزمون را از پس‌آزمون جدا کرد. پیش‌بینی‌های منتج از فرایند (که نمی‌توانسته تحت‌تأثیر تماشای خیابان کنجد باشد، چون به محاسبه پیش‌آزمون محدود شده بود)، در مرحله بعدی با نتایج بدست‌آمده از پس‌آزمون (که احتمالاً تحت تأثیر برنامه «خیابان کنجد» بود) مقایسه شد. معادله رگرسیونی تولیدکننده تخمین رشد ماهانه می‌توانست دربرگیرنده مقیاس محاسبه دیگر تهدیدات روایی (مواردی مانند موقعیت اقتصادی-اجتماعی والدین، یا دیگر محاسبات مرتبط با سوگیری انتخاب) نیز باشد. اگرچه این رویکرد فی‌النفسه دارای مشکلات بسیار جدی است. بدون دانش کامل نسبت به تمام تهدیدات روایی، به ندرت می‌توان از روی نمرات پیش‌بینی‌شده استنباطهای نقیض[۲۹۲] واقعی به دست آورد. بعلاوه، این تحلیل وابسته به تخمینهای پایدار مبتنی بر مقیاسها و اندازه‌گیریهای قابل اتکاء و نمونه‌های بزرگ است. همچنین، این رویکرد نمی‌تواند مشکل امکان وجود اثر گذر زمان در ایجاد اثرات جعلی، و غیرواقعی پس از پیش‌آزمون را مرتفع سازد. دیگر آنکه، احتمال وجود مصنوعات آزمونی وجود دارد، چون نمرات بدست آمده از پس‌آزمون، بر اساس آزمون دوم است، در حالی که اولی (نمرات پیش‌بینی‌شده برای پس‌آزمون) بر اساس آزمون نبود. در نهایت، این شکل از تحلیل اغلب با داده‌های تجمعی در سطح مدرسه[۲۹۳] بکار گرفته می‌شود تا بررسی شود که آیا عملکرد مدرسه در یک سال، از میزان پیش‌بینی شده بهتر بوده است یا نه (پیش‌بینی بر اساس عملکرد آکادمیک قبلی، ماهیت بدنه دانش‌آموزی و شاید حتی بدنه کارمندی و معلمان). به منظور انجام درست این کار، لازم است تا از رویکردهای چندسطحی تحلیل داده استفاده کنیم (Raudenbush & Willms, 1995; Willms, 1992). در این روشها این مسأله که پاسخ‌های فردی درون مدارس لانه گزیده‌اند[۲۹۴]، لحاظ می‌شود. از نگاه نگارندگان، چندبعدی‌سازی رگرسیونی تنها زمانی ارزش انجام دادن دارد که شکل دیگری از گروههای کنترل امکانپذیر نباشد، یا اینکه این روش به عنوان طرحی فرعی برای یک طرح بزرگتر بکار گرفته شود. یقیناً کوک و

292 Counterfactual
293 Aggregated school data
294 Nested

همکارانش این روش را به عنوان تنها یکی از روشهای متعدد یافتن فرضیات در مورد اثربخشی برنامه «خیابان کنجد» بکار گرفتند، و برخی دیگر روشها بسیار متداولتر و متعارفتر هستند.

متضادهای مقایسه‌ای نرُمال شده

در این مورد، عملکرد بدست‌آمده از گروه آزمون در مرحله پیش‌آزمون و پس‌آزمون، با هرگونه نرُم موجود که می‌توانسته برای یافتن استنباط نقیض در زمینه موردنظر مفید باشد، مقایسه می‌شود. برای مثال، ژاکوبسن و همکارانش (Jacobson et al., 1984) زوجهای تحت‌درمان برای مشکلات زناشویی را با نرُمهای موجود در مورد زوجهای متعادل در مقیاس سازگاری ازدواج مقایسه کردند، تا ببینند که آیا خانواده‌درمانی می‌تواند زوجها را سازگارتر نماید؟ به همین ترتیب، نیتزل و همکارانش (Nietzel et al., 1987) مطالعاتی را مورد بررسی قرار دادند که در آنها گروههای دریافت‌کننده درمان افسردگی را با نرُم مقیاس افسردگی بک۲۹۵ مقایسه می‌کردند (Beck, Ward, Mendelsohn, Mock, & Erbough, 1961)، تا ببینند که آیا این افراد به سطوح حال خوب روانی همتراز با میزان گزارش‌شده برای افراد غیرافسرده رسیده‌اند؟ در هر یک از این مطالعات، نمرات پس‌آزمونی گروه آزمون، با نرُمها مقایسه می‌شوند. اگر گروه آزمون توانسته بود در سطح یا بالاتر از سطح استانداردهای نُرمی مربوطه عمل نماید، اثرات از نظر کلینیکی معنادار شناخته می‌شدند.

این شکل از مقایسه در مطالعات تحصیلی، و برای بررسی اینکه آیا یک گروه از دانش‌آموزان، کلاسها، یا مدارس توانسته‌اند در طول زمان از نظر رتبه‌بندیهای منتشرشده پیشرفت نمایند، مورد استفاده قرار می‌گیرد.

رتبه‌بندی‌ها از روی نُرمهای منتشر شده اقتباس می‌شوند، و با هدف نشان‌دادن عملکرد دانش‌آموزان، و تغییرات آن در مقایسه با عملکرد دانش‌آموزان مورد بررسی در نمونه نُرمال‌سازی اصلی، بکار می‌روند. متضاد هنجاری یک نقیض بسیار ضعیف است، و به سختی می‌تواند مشخص کند که چطور رفتار شرکت‌کنندگانی که بواقع تحت مداخله قرار گرفته‌اند، می‌توانست در غیاب درمان متفاوت باشد. در واقع تا آن زمان که گروه هنجاری به گونه‌ای انتخاب می‌شود که برتر از گروه تحت مداخله عمل کند، غیرممکن است از مقایسه میان آنها بتوان اثر مداخله را بدست آورد، حتی اگر مداخله نتایج را، در مقایسه با حالت نبود مداخله، به میزان قابل‌توجهی افزایش داده باشد. حتی مداخلههایی که در مقایسه با یک گروه کنترل استاندارد از سطح بالایی از کارایی برخوردار هستند، در قیاس با چنین نُرمهایی غیرکارآمد قلمداد می‌شوند. سوگیریهای مختلفی می‌توانند در جریان مقایسه هنجاری تهدیدآفرین باشند. اول سوگیری انتخاب، زیرا نمونه نُرم‌شده معمولاً با نمونه دریافت‌کننده مداخله تفاوت دارد؛ دوم، بواسطه خطر گذشت زمان، چون نُرمها معمولاً به زمانی پیش از زمان انجام مداخله روی گروه آزمون تعلق دارند. سوم، بواسطه خطای آزمون، اگر روی نمونه تحت مداخله پیش‌آزمون انجام شده باشد، اما روی نمونه

295 Beck depression inventory

نُرم‌شده انجام نشده باشد. چهارم، بواسطه خطر رگرسیون، اگر نمونه تحت مداخله به دلیل نیاز زیاد (یا شواهد زیاد) انتخاب شده باشد، اما نمونه نُرم‌شده اینطور نباشد. پنجم، بواسطه خطر ابزار، اگر شرایط اندازه‌گیری تا حد زیادی برای گروههای مداخله و نُرم‌شده متفاوت باشد. و در نهایت، بواسطه خطر بلوغ، اگر گروه تحت مداخله در قیاس با گروههای نُرم‌شده به سرعت تغییر می‌کند. برخی اوقات می‌توان این خطرات را از طریق روشهایی مانند استفاده از نُرمهای محلی جمع‌آوری شده از همان جامعه (که گروه مداخله از آن گرفته می‌شود)، اطمینان حاصل کردن از یکسان بودن شرایط و زمانبندی یکسان برای گروهها، و همچنین انتخاب نمونه‌های هنجاری که احتمال می‌رود تجربیات بلوغ مشابه داشته باشند، تقلیل داد.

متضادهایی با منبع ثانویه[۲۹۶]

حتی بدون نُرمهای منتشر شده نیز، محققین می‌توانند متضادهایی را با استفاده از منابع ثانویه بسازند. برای مثال، محققین پزشکی برخی مواقع از مواردی نظیر سریهای کلینیکی، و مستندات مربوط به مراجعین درمان‌شده قبل از معرفی درمان جدید، برای این منظور بهره می‌گیرند. اطلاعات درمانی ثبت‌شده تمامی بیماران با یک شرایط خاص، مورد استفاده قرار می‌گیرد، و برخی اوقات نیز کنترلهای انجام‌شده در طول زمان، روی افراد داخل همان مؤسسه، در دسترس است. مطالعات انجام‌شده روی اثر برنامه‌های پیش‌گیری از سوءمصرف مواد، نتایج طرحهای آزمایش تک‌گروهی خود را با کمک داده‌های بدست‌آمده از پیمایشهای ملّی و ایالتی تکمیل کردند (Furlong, Casas, Corral, & Gordon, 1997; Shaw, Rosati, Salzman, Cole, & McGeary, 1997).

اقتصاددانان اشتغال، اطلاعات جمع‌آوری شده در سطح ملّی را به همین شیوه به کار می‌گیرند، و بر اساس پیمایشهای جمعیتی فعلی، یا مطالعات پانل در مورد تحرکات درآمدی، متضادهایی می‌سازند که با استفاده از آن برنامه‌های آموزش حین‌خدمت را ارزیابی می‌کنند. این متضادها به عنوان شاخص اولیه برای تشخیص امکان وجود اثر مداخله مفید واقع می‌شوند (برای مثال، در فاز دوم درمان در پزشکی، برای اینکه ببینیم درمان جدید قابل‌اعتماد است یا نه).

اگرچه هرگونه استفاده از چنین داده‌های مربوط به گذشته و آرشیوی، با مشکلات عملی جدی شبیه همان مشکلاتی که در مورد نمونه‌های نُرم‌شده ذکر شد مواجه است. اطلاعات ممکن است برای اهدافی متفاوت از اهداف مطالعه حاضر جمع‌آوری شده باشد. این باعث می‌شود تا علیرغم شباهتهای سطحی و ظاهری در توضیحات منتشرشده، قیاس‌پذیری دو نمونه کاهش بیابد. ممکن است داده کیفیت ضعیفی داشته باشد، مثلاً آزمونهای پایایی انجام‌شده روی داده‌های مداخله ممکن است روی داده‌های آرشیوی انجام نشده باشد، و در این

296 Secondary source contrasts

داده‌ها به طور معمول داده‌های گم‌شده[۲۹۷] وجود دارد. داده‌ها ممکن است فاقد متغیرهای کمکی[۲۹۸] موردنیاز برای تعدیل تفاوتهای میان گروهها و آسیب‌شناسی قیاس‌پذیری گروهها باشد. این متضادها می‌توانند سوگیری قابل توجهی در جریان تخمین اثرات ایجاد کنند (Sacks, Chalmers, & Smith, 1982, 1983). این سوگیریها می تواند به دلیل تغییر در جمعیهایی که مبتلا به مشکل هستند (مثلاً جمعیت مبتلایان به ایدز طی زمان دستخوش تغییر شده است)، و یا تغییر در جمعیتی که واجدشرایط دسترسی به درمان بوده، و یا در حال حاضر به درمان دسترسی دارند (مثلاً تغییراتی که بواسطه آن، مداخلههای مقابله با ایدز قابل‌پیگیری می‌شوند) به مطالعه تحمیل شوند. بنابراین باز به این مسأله برمی‌گردیم که استفاده از این منابع ثانویه متضادها هنگامی بیشترین منفعت را برای اهداف تحقیق دربرخواهد داشت، که با دیگر عناصر طرح تلفیق بشود، و شبه‌آزمایش پیچیده‌تری بوجود بیاورد.

طرحهای مورد-کنترل[۲۹۹]

در طرحهایی که تا به حال مورد بررسی قرار دادیم، شرکت‌کنندگان به گروههایی تقسیم می‌شوند که یک مداخله را دریافت می‌کنند یا نمی‌کنند، و سپس نتایج آنها مورد بررسی قرار می‌گیرد. این تلاش برای یافتن اثرات علّتها، مشخصه آزمایشهاست. اگرچه برخی اوقات انجام آزمایش اخلاقی نیست، و یا توجیه اقتصادی ندارد. به عنوان نمونه، در سال ۱۹۶۰ داروی دیتیل‌استیل‌سیترول(DES) برای زنان بارداری که بواسطه خونریزی در خطر سقط جنین بودند، تجویز شد. پس از مدتی این شائبه بوجود آمد که این دارو می‌تواند منجر به ایجاد سرطان واژن در دختران زنانی شود که دارو را مصرف کرده‌اند. این که دانسته این دارو را برای برخی تجویز، و برای برخی دیگر تجویز نکنیم، غیراخلاقی بود. بعلاوه، سرطان واژن بسیار نادر است، و زمان بسیار طولانی طول می‌کشد تا شکل بگیرد؛ بنابراین، اندازه نمونه بسیار بزرگ و سالهای متمادی زمان لازم بود تا بتوان نتایج آزمایشی با روایی قابل‌قبول بدست آورد. در این موارد، راه‌حل جایگزین، استفاده از طرح مورد-کنترل است (این طرح همینطور با نامهای مقایسه موارد، تاریخ-موارد و طرحهای پس‌نگر نیز شناخته می‌شود) که در تحقیقات اپیدمی‌شناسی پایه‌گذاری شده، و رواج قابل‌توجهی دارد. در این طرحها، یک گروه، متشکل از افرادی است که نتایج موردنظر محقق را دارا هستند، و گروه دیگر افراد مربوط به گروه کنترل که فاقد نتایج مورد نظر هستند را در برمی‌گیرد. نتایج در این طرحها غالباً در مقیاس دوتایی بدست می‌آیند؛ مانند رد یا قبول، سالم یا بیمار، زنده یا مرده، متأهل یا مجرد، سیگاری یا غیرسیگاری، پاکی یا مصرف دوباره مواد مخدر، افسرده یا غیرافسرده، رشدکرده یا نکرده. موارد و کنترلها سپس با استفاده از داده‌های پس‌نگر مورد مقایسه قرار می‌گیرند تا ببینیم که

297 missing
298 Covariate
299 Case-control

آیا موراد نسبت به کنترلها به میزان بیشتری در معرض علّتهای مفروض بوده اند؟ هربست و همکارانش (Herbst et al., 1971) هشت مورد مبتلا به سرطان واژن را یافتند و آنها را با سی و دو نفر کنترل که سرطان واژن نداشتند، جفت کردند. همه این افراد طی پنج روز در همان بیمارستان بدنیا آمده بودند. هفت نفر از هشت مورد، DES دریافت کرده بودند اما هیچ کدام از کنترلها DES مصرف نکرده بودند.

طرح مورد-کنترل، طرحی عالی برای تدوین فرضیاتی در خصوص روابط علّی است. از جمله روابط علّی‌ای که با استفاده از مطالعات مورد-کنترل بدست آمده‌اند، می‌توان به رابطه میان سیگار کشیدن و سرطان، قرصهای ضدبارداری، و گرفتگی عروق، و DES و سرطان واژن اشاره کرد (Vessey, 1979). مطالعات مورد-کنترل در مواقعی که نتایج موردنظر نایاب بوده، و سالها طول می‌کشد تا بوجود بیایند یا رشد کنند، بسیار توجیه‌پذیر است. این طرحها غالباً ارزانتر و از نظر لجستیکی آسانتر هستند، مخاطرات غیرضروری متوجّه شرکت‌کنندگان، که می‌تواند در اثر مداخلههای آزمایشی ایجاد شود را کاهش می‌دهند، و امکان آزمون علّتهای متعدد برای یک شرایط خاص را فراهم می‌آورند (Baker & Curbow, 1991).

برخی مشکلات روش‌شناختی در مطالعات مورد-کنترل مرسوم هستند. تعریف و انتخاب موردها نیازمند تصمیم‌گیری دقیق درباره آن چیزی است که تعیین‌کننده وجود یا عدم وجود نتیجه‌ی مورد نظر محقق است. دیگر محققین فعال در زمینه تحقیق موردنظر ممکن است با تعریف محقق موافق نباشند، و حتی اگر موافق باشند، روشهای ارزیابی نتایج ممکن است غیرقابل‌اتکاء بوده، و روایی پایینی داشته باشند. بعلاوه، تعاریف و مقیاسها ممکن است در طول زمان تغییر کنند، و بنابراین، برای افرادی که اخیراً به عنوان بیمار تشخیص داده شده‌اند، متفاوت باشد. حتی زمانی که تمامی موارد در بازه زمانی نزدیک به هم تشخیص داده شده باشند، این افراد غالباً به این دلیل مشخص شده‌اند که نتایج درمانی آنها، ایشان را مورد توجه محقق قرار داده است؛ مانند هنگامی که زنان با علائم سرطان واژن، برای درمان و تشخیص به محقق ارایه کردند. کنترلها به ندرت به این شیوه به محقق مراجعه می‌کنند، زیرا آنها دچار این علائم نیستند. به همین دلیل، مکانیزمهای انتخاب دو گروه به طور اجتناب‌ناپذیری متفاوت است. ریزش از دیگر مشکلات است. علّتهای موردنظر محقق می‌تواند موجب ریزش در نمونه شوند (مانند هنگامی که برخی افراد قبل از شروع مطالعه، به دلیل سرطان می‌میرند). افراد حذف‌شده به این دلیل، ممکن است دارای مشخصاتی متفاوت از افراد باقی‌مانده در مطالعه بوده باشند، تفاوتی که اگر توزیع در کنترلها به شیوه‌ای مشابه تغییر نکرده باشد، می‌تواند موجب سوگیری شود.

انتخاب موردها در طرح مورد-کنترل دشوار اشت. یک روش معمول برای انتخاب کنترلها آن است که کنترلها را به گونه‌ای انتخاب کنیم که نماینده جمعیت موردنظر باشند. کنترلهایی که به صورت تصادفی نمونه‌گیری شده‌اند، مثال خوبی از این روش است. اما هنگامی که نمونه‌گیری تصادفی مقرون‌به‌صرفه نباشد، کنترلها معمولاً از طریق جفت‌کردن کنترلها با موارد، از نظر مشخصاتی که می‌تواند با نتایج ارتباط داشته باشد، انتخاب می

شوند. انتخاب کنترلهای همسایه (یعنی افرادی که در معرض محیط زندگی مشابهی هستند)، و کنترلهای در حال گذران درمان در بیمارستان مشترک (که تحت‌درمان با همان امکانات درمانی هستند)، از انواع معمول کنترلهای جفت‌شده به حساب می‌آیند (Lund, 1989). روش هربست و همکارانش (Herbst et al., 1971) برای انتخاب کنترلها، از میان آنها که در همان بیمارستانی بدنیا آمده بودند که افراد مورد بدنیا آمده بودند، احتمالاً می‌توانست شباهت میان کنترلها و موارد را از نظر اثرات جغرافیایی، جمعیت‌شناختی، و اثرات همگروههای دوران تولد[۳۰۰] افزایش دهد. اگرچه، کنترلهای جفت‌شده همچنان از بسیاری جهات مشاهده‌نشده متفاوت خواهند بود؛ عواملی که می‌توانند به عنوان متغیر مزاحم با اثر علّت مخلوط شوند، و یا حتی می‌توانند خود علّت واقعی بروز نتایج باشند. برای مثال، در یک مطالعه بر روی کودکان مبتلا به دیابت، «کنترلهای دوستان» به کار گرفته شدند، که در آن از والدین خواسته می‌شد که نام دو نفر از دوستان کودک که از نظر جنسیت و سن با کودک همسان بودند را به عنوان کنترل معرفی کنند (Siemiatycki, Colle, Campbell, Deware, & Belmonte, 1989). نتایج نشان داد که کودکان مبتلا به دیابت در مقایسه با دیگر کودکان، به احتمال بیشتر در مدرسه دچار مشکل هستند، دوستان کمتری دارند، برای خوابیدن مشکل دارند، بستری شده‌اند و دچار تصادف شده‌اند، و والدینشان از یکدیگر جدا شده‌اند. اما این متغیرها علّت هستند یا متغیر کمکی؟ بررسی‌های بعدی مشخص کرد که والدین افراد اجتماعی‌تر را به عنوان دوستان کودکان خود معرفی کرده بودند. بنابراین، گروه کنترل به طور نامتوازن با گروه موردها از نظر خصوصیات اجتماعی مثبت متفاوت بودند (Siemiatycki, 1989). بکارگیری کنترلهای متعدد می‌توانست از بروز این مشکل جلوگیری کند (Kleinbaum, Kupper, & Morgenstern, 1982). مثلاً انتخاب گروهی از همان درمانگاه که کودکان دیابتی به آن مراجعه می‌کردند، انتخاب گروه دیگری از همسایه‌های این کودکان، و در نهایت، انتخاب نمونه‌ای تصادفی از میان کل جمعیت کودکان (Baker & Curbow, 1991; Lund, 1989). تفاوتهای موجود میان کنترلها از نظر تخمینهای روابط علّی، کمک می‌کند تا میزان سوگیریهای پنهان ممکن را به نحوه بهتری شناسایی کنیم.

بعلاوه، این که چه جمعیت کنترلی مرتبط‌ترین به گروه موردهاست، تا حد زیادی به استنباط مورد نظر بستگی دارد. برای مثال، همسایه‌های یک مورد عموماً به عنوان کنترل بکار گرفته می‌شوند. اما اگر سؤال این باشد که اسهال مشاهده‌شده در مسافران آمریکایی، در یک بیمارستان مکزیکی، به دلیل نوشیدن تکیلا ایجاد شده است یا خیر، انتخاب همسایگان چندان مناسب نخواهد بود. در عوض، می‌توان دیگر مسافران غیرمکزیکی مراجعه کرده به همان بیمارستان در مکزیک را به عنوان گروه کنترل مورد استفاده قرار داد (Miettinen, 1985). گروه کنترل انتخاب‌شده از میان کل جمعیت، برای زمانهایی مناسب است که دانش اندکی در مورد علل خاص در دست داریم. اما گروههای کنترل که دقیقاً تعریف می‌شوند، در شرایطی که سوال علّی دقیقاً مشخص است کارآمدتر

300 Birth cohort effect

خواهند بود (Garber & Hollon, 1991). همچنین، مواردی که دچار یک مشکل مشترک هستند، الزاماً از نظر علّت بروز مشکل شبیه و مشترک نیستند. برای مثال، اگر موارد بیماران مبتلا به عفونتهای استافیلوکوکی در یک بیمارستان هستند، برخی ممکن است بیرون از بیمارستان دچار عفونت شده‌اند، و برخی داخل بیمارستان (عفونتهای ناشی از جراحی یا درمان پزشکی)، که این دو گروه نیازمند گروههای کنترل متفاوتی هستند.

ارزیابی دریافت مداخله در مطالعات مورد-کنترل بصورت پس‌نگر، و با استفاده از منابع مبتلا به خطا و سوگیری، مانند حافظه یا رکوردهای سازمانی صورت می‌گیرد. از این رو، طبقه‌بندی افراد به دو گروه در دریافت کننده مداخله و کنترل به ندرت به طور کامل و صحیح انجام خواهد شد. موارد در مقایسه با کنترلها انگیزه قویتری برای به خاطر آوردن موقعیتهایی که در معرض عوامل خطرزا قرار گرفته‌اند، دارند. برای مثال، آنها ممکن است چنین برداشت کنند که دقت و صحت درمان آنها در گرو دادن اطلاعات دقیق‌تر است. علاوه بر این، بودن در معرض مداخله مطمئناً با اثر دیگر متغیرهای کمکی مخلوط شده است. در مثال DES، به مادران DES داده شده بود چون در معرض خطر سقط جنین قرار داشتند. بنابراین ریسک سقط خودبخودی ناشی از خونریزی می‌توانسته با اثر DES مخلوط شده باشد. البته در این مثال، اثر این ریسک بر سرطان واژن غیرموجه بوده است، چون نتایج دیگر تحقیقات همبستگی تأییدکننده وجود رابطه میان DES و سرطان واژن است، آزمایشات تصادفی روی حیوانات تأییدکننده اثر است، و همچنین زمینه کافی برای تأیید رابطه میان سقط جنین و سرطان واژن وجود دارد. با این وجود اثر بودن در معرض مداخله، عموماً با اثر دیگر متغیرهای مزاحم مخلوط می‌شود، و بنابراین برقرارای رابطه روشن میان مداخله و نتایج را هر چه بیشتر دشوار می‌سازد.

از این مثالها می‌توان دریافت از آنجا که مطالعه مورد-کنترل به دنبال استنباطهای علّی است، منطق بی‌اثرسازی تهدیدات روایی درونی برای آن مصداق دارد. در واقع ادبیاتی مستقل از آنچه در این کتاب درباره روایی ارایه شد، در مورد تهدیدات روایی در مطالعات مورد-کنترل وجود دارد. استنباطهای علّی بدست آمده از مطالعات مورد-کنترل وجود دارد علی‌الخصوص برای تعدیل اثرات ناشی از متغیرهای مزاحم. با این وجود، نگارندگان بر این باورند که طرحهای مورد-کنترل قابلیت بکار گرفته شدن در زمینه‌هایی غیر از سلامت عمومی را نیز دارند، این قابلیت اگرچه بیشتر در تولید فرضیات علّی است تا در آزمون آن فرضیات.

فصل ۵

طرحهای شبه‌آزمایشی دارای گروه کنترل و پیش‌آزمون

کنترل: به عنوان فعل: ۱.الف. تأیید یا تنظیم کردن (یک آزمایش علّی) از طریق انجام آزمایشی موازی یا بوسیله مقایسه با استانداردی دیگر. ب. تأیید (مثلاً یک حساب) با استفاده از ثبت دوباره (یا دوتایی) برای مقایسه. به عنوان اسم: ۱.الف. یک استاندارد مقایسه برای بررسی یا تأیید نتایج یک آزمایش. ب. فرد یا گروهی که به عنوان استاندارد مقایسه در یک آزمایش کنترل مورد استفاده قرار می‌گیرد.

پیش‌آزمون: ۱.الف. آزمونی اولیه که برای تعیین اینکه آیا دانش‌آموزان به اندازه کافی برای مطالعات پیشرفته بعدی آماده شده‌اند، انجام می‌شود. ب. آزمونی که به منظور تمرین انجام می‌شود. ۲. آزمون پیشرفته، چیزی شبیه یک پرسشنامه، یک محصول و یا یک ایده.

طی برنامه‌ای با عنوان خانه‌دار، به تعداد انتخاب شده‌ای از افراد یارانه‌بگیر برای مدت شش هفته آموزش ارائه شده، و به دنبال آن، ایشان را برای شغل خانه‌داری و پیشکاری در خانه به کار گرفتند. برای بررسی اینکه آیا آموزش ارائه شده باعث افزایش درآمدهای بعدی ناشی از شغل شده است یا نه، بل و همکارانش (Bell et al., 1995) نتایج بدست آمده از این گروه را با سه گروه کنترل غیرتصادفی مقایسه نمودند: (۱) افرادی که برای برنامه درخواست داده بودند، اما قبل از آنکه صلاحیت آنها برای برنامه بررسی شود، آن را ترک کرده بودند؛ (۲) افرادی که توسط مجریان برنامه واجد شرایط تشخیص داده نشده بودند؛ (۳) افرادی که ثبت‌نام کرده و واجد شرایط تشخیص داده شده بودند، اما در آموزش شرکت نکردند. مقایسه نتایج گروه آزمون با سه گروه کنترل مذکور نشان می‌داد که آموزشهای ارائه‌شده درآمدهای بعدی را افزایش داده، اما اندازه اثر بسته به نوع گروه کنترل متفاوت بوده است. اطلاعات درآمدی هر سه گروه کنترل قبل از انجام مداخله موجود بود، و بل و همکارانش (۱۹۹۵) نشان دادند که احتمال کمی وجود دارد که تفاوتهای پیش‌آزمون موجب تفاوتهای مشاهده‌شده در پس‌آزمون شده باشد.

طرحهایی که گروه کنترل و پیش‌آزمون دارند

این فصل بر طرحهایی تمرکز داردکه مانند مطالعه بل و همکارانش، هم گروه کنترل دارند و هم پیش‌آزمون. فصل حاضر نشان می‌دهد که انتخاب دقیق گروههای کنترل چطور می‌تواند استنباطهای علّی بدست‌آمده از آزمایشها را تسهیل نماید؛ اما ضمناً نشان می‌دهد که داشتن چنین گروههای کنترلی فایده اندکی دارد، مگر آنکه با پیش‌آزمونی همراه باشد که همان متغیرهایی را مورد بررسی قرار می‌دهد که در پس‌آزمون اندازه‌گیری می‌شود. چنین پیش‌آزمونهایی فواید فراوانی دارند. این پیش‌آزمونها باعث می‌شوند تا متوجه شویم گروههای مورد مقایسه از ابتدا متفاوت بوده‌اند، و بنابراین هوشیاری محقق نسبت به وجود نوعی تهدید روایی درونی (و نه

دیگر متغیرهای مستقل) اثرگذار افزایش می‌یابد. آنها همچنین می‌توانند کمک کنند تا بزرگی تفاوتهای اولیه میان گروهها را از نظر متغیری که معمولاً بیشترین همبستگی را با نتایج دارد، ارزیابی کنیم. فرضیه قوی آن است که هر چه تفاوت مشاهده شده در پیش‌آزمون کوچکتر باشد، احتمال آنکه سوگیری انتخاب قابل‌توجهی در آن پیش‌آزمون دخیل بوده باشد، کمتر خواهد بود. اگرچه، بر خلاف تخصیص تصادفی، هیچ تخصیصی وجود ندارد که در آن بتوان فرض کرد متغیرهای محاسبه نشده پیش‌آزمون بی ارتباط به متغیر نتیجه‌ای هستند. در نهایت، داشتن مقیاسهای پیش‌آزمون کمک فراوانی در تحلیلهای آماری خواهد کرد، علی‌الخصوص اگر پایایی این مقیاسها از پیش معلوم باشد. هیچ متغیر دیگری نمی‌تواند به خوبی پیش‌آزمون این اهداف را برآورده نماید.
تمامی این دلایل روشن می‌کند که چرا در این فصل به مرور پیش‌آزمونها و گروههای کنترل در شبه‌آزمایشها خواهیم پرداخت. جدول ۵.۱ خلاصه‌ای از طرحهای شبه‌آزمایشی موردنظر را نشان می‌دهد.

جدول ۵.۱ طرحهای شبه‌آزمایشی که پیش‌آزمون و گروه کنترل دارند

طرح گروه کنترل دستکاری نشده با نمونه‌های پیش‌آزمون و پس‌آزمون وابسته

NR O_1 X O_2

NR O_1 O_2

طرح گروه کنترل دستکاری نشده با نمونه‌های پیش‌آزمون و پس‌آزمون وابسته با دو پیش‌آزمون

NR O_1 O_2 X O_3

NR O_1 O_2 O_3

طرح گروه کنترل دستکاری‌نشده با نمونه‌های پیش‌آزمون و پس‌آزمون وابسته، با تکرارهای جابجا شونده[۳۰۱]

NR O_1 X O_2 O_3

NR O_1 O_2 X O_3

طرح گروه کنترل دستکاری‌نشده با نمونه‌های پیش‌آزمون و پس‌آزمون وابسته، با گروه کنترل معکوس دستکاری‌شده

NR O_1 X_+ O_2

NR O_1 X_- O_2

301 Switching replication

طرح گروه کنترل همتایان[۳۰۲]

NR O_1

NR X O_2

طرح گروه کنترل همتایان با پیش‌آزمون از هر همتا

NR O_1 O_2

NR O_3 X O_4

طرح گروه کنترل دستکاری‌نشده، با نمونه‌های پیش‌آزمون و پس‌آزمون وابسته

این طرح که غالباً از آن با عنوان طرح گروههای مقایسه غیرهم‌ارز یاد می‌شود، را می‌توان معمول‌ترین نوع شبه‌آزمایشها دانست. طرح اولیه‌ای (واریته اول از این نوع طرحها) که مورد بررسی قرار می‌دهیم، یک گروه مداخله و یک گروه کنترل دستکاری‌نشده را بکار می‌گیرد، و در آن، داده‌های پیش‌آزمون و پس‌آزمون از افرادی یکسان جمع‌آوری می‌شود[۳۰۳]؛ که این همان مشخصه *نمونه‌های وابسته* است. این طرح را به صورت زیر نشان می‌دهند:

NR O_1 X O_2

NR O_1 O_2

استفاده همزمان از پیش‌آزمون و یک گروه کنترل آزمون برخی تهدیدات روایی را راحت‌تر می‌کند. از آنجا که گروهها به لحاظ تعریف غیرهم‌ارز هستند، وجود سوگیری انتخاب امری مفروض است. پیش‌آزمون امکان کشف اندازه و جهت این سوگیریها را میسر می‌نماید. به عنوان نمونه، کارتر و همکارانش (Carter et al., 1987) در

---

302 Cohort

۳۰۳ یک واریته (variation) طرح جابجایی نقطه رگرسیونی است. این طرح پس‌آزمون، یک متغیر پیش‌بین از نمرات پس‌آزمون که قبل از مداخله جمع آوری می شود (پیش‌بین می تواند پیش‌آزمون باشد اما غالباً نیست)، و یک واحد مداخله و تعداد زیادی واحد کنترل (از هر واحد یک میانگین گروهی، و نه اطلاعات مربوط به افراد داخل واحدها، در آنالیزها وارد می شود)(Campbell & Russo, 1999). این طرح زمانی که یک پیش‌آزمون منفرد (یا دیگر پیش‌بینها) و پس‌آزمون از تعداد بسیار محدودی واحد مداخله در اختیار داریم و هیچ طرح دیگری شدنی و اقتصادی نیست، می تواند سودمند باشد. این حالت می تواند در مواردی که اسناد اجرایی به صورت تجمعی ثبت شده اند و تعداد خیلی زیادی کنترل وجود دارد، و یا در زمینه‌های کلینیکی که در آنها یک درمان خاص به تنها یک مراجع خاص داده می‌شود اما اطلاعات مربوط به تعداد زیادی کنترل در دسترس است، رخ دهد.

مطالعه‌ای، اثر جایزه پیشرفت پژوهشی سازمان بهداشت ملی (برنامه‌ای که برای ارتقاء مشاغل پژوهشی دانشمندان برجسته طراحی شده بود) را مورد بررسی قرار دادند. آنها دریافتند که کسانی که این جایزه را دریافت کرده بودند، عملکرد بهتری نسبت به کسانی که در یافت نکرده بودند داشتند، اما این افراد در پیش‌آزمون نیز به همین میزان عملکرد بهتری داشتند. در نتیجه، اثر مشاهده شده احتمالاً بیشتر به دلیل سوگیری انتخاب بوده است تا اثر جایزه مورد بحث. استفاده از پیش‌آزمون امکان شناسایی ماهیت ریزشها را نیز فراهم می‌آورد، و به محققین اجازه می‌دهد بتوانند تفاوت میان افرادی که در مطالعه می‌مانند، و آنها که مطالعه را ترک می‌کنند را توضیح دهند. اگرچه اینکه پیش‌آزمون تا چه اندازه بتواند خطر سوگیری انتخاب را برطرف کند، به اندازه سوگیری موردنظر، و نقش هر متغیر محاسبه‌نشده‌ای که موجب انتخاب شده، و با نتایج همبستگی دارد وابسته است. البته نبود تفاوت در مرحله پیش‌آزمون در یک شبه‌آزمایش به هیچ عنوان به معنی عدم‌وجود سوگیری انتخاب نیست.

وقتی گروهها در پیش‌آزمون تفاوت داشته باشند، این امکان که سوگیری انتخاب به صورت برهم‌کنشی (تعاملی) و یا تجمعی با دیگر تهدیدات روایی ترکیب شود وجود دارد. برای مثال، اگر پاسخ‌دهندگان در یک گروه، در مقایسه با گروه دیگر، به میزان بیشتری با تجربه، خسته و یا بی‌حوصله شوند، *تهدید ترکیبی انتخاب-بلوغ* می‌تواند بروز کند. برای روشن شدن این موضوع، فرض کنید در موقعیتی که میانگین نمرات گروه آزمون در پیش‌آزمون بیشتر از میانگین نمرات گروه کنترل است، روشی جدید معرفی می‌شود. اگر مداخله مورد نظر عملکرد را ارتقاء دهد، تفاوت میان گروهها ممکن است بزرگتر از تفاوت مشاهده شده اولیه (در پیش‌آزمون) باشد. اما الگوی مشاهده شده می‌توانست در حالتی که افراد گروه آزمون باهوشتر از گروه کنترل باشند و میل بیشتر خود برای یادگیری را با سرعتی بیشتر از گروه کنترل بکار گرفته باشند، نیز رخ دهد. یعنی غنی ثروتمندتر شده باشد.

زمانی که گروههای غیرهم‌ارز در پیش‌آزمون از نقاط متفاوتی شروع کنند، *تهدید انتخاب-ابزار* مجال بروز پیدا می‌کند. در بسیاری از مقیاسها، فاصله‌ها (بازه‌ها) نابرابر هستند، و تغییرات را در برخی نقاط مقیاس آسانتر می‌توان تشخیص داد (مثلاً در میانه طیف به نسبت انتهاهای طیف). به عنوان نمونه، در نمرات آزمون نُرم‌شده موفقیت، بدست‌آوردن پاسخ صحیح در یک مقیاس تک‌گویه‌ای کاربرد و مفهوم بیشتری برای رتبه‌بندی درصدی در دو سر توزیع دارد تا در محل میانگین آن. بنابراین بسته به جایگاه پاسخگو روی مقیاس، یک گویه می‌تواند به مقادیر متفاوتی از تغییرات درصدی تفسیر شود. تهدید انتخاب-ابزار احتمالاً زمانی حادتر می‌شود که (۱) غیرهم‌ارزی اولیه بیشتری میان گروهها وجود داشته باشد، (۲) تغییرات پیش‌آزمون-پس‌آزمون بزرگتر باشد، و (۳) میانگین هر گروه به انتهای طیف مقیاس نزدیکتر باشد، و بنابراین اثر سقف یا کف رخ دهد. برخی اوقات برای بررسی احتمال وجود مشکل انتخاب-ابزار، می‌توان توزیع فراوانی‌های درون هر گروه در پیش‌آزمون و

پس‌آزمون را مورد بررسی قرار داده، و چولگی‌های احتمالی و یا همبستگی میان میانگین و واریانس گروهها را شناسایی کرد. بعضاً می‌توانیم داده‌های خام را مقیاس‌گذاری مجدد[۳۰۴] کنیم (مقیاس آنها را تغییر دهیم)، در حالی که در برخی دیگر از اوقات لازم است تا به دقت گروههایی را انتخاب کنیم که نمراتی نزدیک به هم داشته، و در میانه مقیاس قرار می‌گیرند.

سومین نمونه از تهدیدات ترکیبی، تهدید انتخاب-رگرسیون است. در شبه‌آزمایش پیش‌دبستان که در فصل پیش توضیح داده شد (Cicerelli & Associates, 1969)، گروه آزمون کودکانی که در برنامه پیش‌دبستانی شرکت کرده بودند، به طور بالقوه از جمعیت متفاوتی (نسبت به جمعیتی که گروه کنترل از آن گرفته شده بود) انتخاب شده بودند. با درک احتمال وجود چنین مشکلی، محققین در آن پژوهش یک گروه کنترل هم ارز انتخاب کردند؛ و تنها داده‌های مربوط به کسانی را از گروه کنترل در نظر گرفتند که از نظر جنسیت، سن، و وضعیت حضور در مهدکودک مشابه کودکان گروه آزمون بودند. چهارمین مشکل، تهدید ترکیبی انتخاب-گذشت زمان[۳۰۵] است (یا گذشت زمان محلی[۳۰۶]). این احتمال که در فاصله زمانی میان پیش‌آزمون و پس‌آزمون رویدادهایی رخ داده باشد که بر یک گروه بیشتر اثر گذاشته باشد. برای مثال، نتایج مطالعه‌ای که به مرور دستاوردهای برنامه‌های دولت فدرال برای ارتقاء باردارها می‌پرداخت (Shadish & Reis, 1984)، نشان داد که مطالعات متعددی طرحهای پیش‌آزمون-پس‌آزمون را مورد استفاده قرار داده، و نتایج آنها حاکی از اثربخش بودن برنامه مذکور در ارتقاء نتایج بارداریها بوده است. اما مادرانی که واجد شرایط شمول در این برنامه بودند واجد شرایط شمول در دیگر برنامه‌هایی که می‌توانسته باعث بهبود بارداری شوند (مانند برنامه تغذیه مادر و کودک و دیگر برنامه‌های بهداشتی) نیز بودند. بنابراین امکان نداشت بتوانیم با اطمینان بگوییم که بهبود مشاهده‌شده در بارداریها به علّت مداخله بکار گرفته شده در برنامه پژوهش (و یا در اثر دیگر برنامه‌ها) ایجاد شده است.

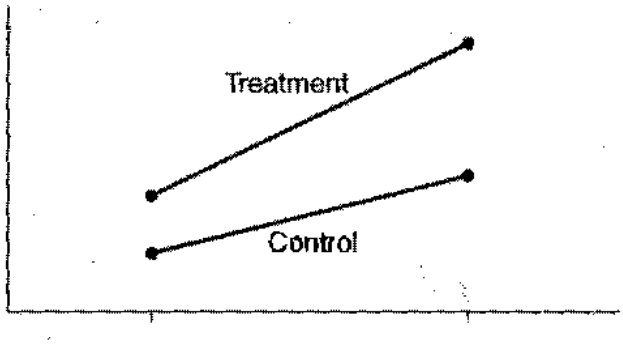
موجّه‌بودن تهدیدات تا حدی وابسته به الگوی نتایج مشاهده شده است

فهرست تهدیدات روایی دورنی مرتبط با این طرحها دلهره‌آور و مضطرب‌کننده است. با این وجود، موجّه‌بودن یک تهدید تا حد زیادی به زمینه تحقیق بستگی دارد[۳۰۷]. به بیان دقیق‌تر، به مشخصات مشترک طرح آزمایش، دانش موجود نسبت به تهدیدات، و الگوی نتایج مشاهده‌شده مطالعه وابسته است. در نتیجه، همه تهدیدات روایی همواره موجّه نیستند. مثلاً، فرایندهای بلوغ در کودکان که باعث افزایش پیشرفت تحصیلی دانشجویان شده، توضیح موجّهی برای کاهش مشاهده شده در پیشرفت تحصیلی نیست. برای آنکه به طور کلی‌تر به این موضوع بپردازیم، در این قسمت پنج الگوی نتایج که در طرحهای مقایسه گروهی پیش‌آزمون-پس‌آزمون مشاهده

304 Rescale
305 History
306 Local history
307 Context-dependent

می‌شود را ارائه نموده، و نشان خواهیم داد که این طرحها چطور می‌توانند خطرات کم و بیش موجّهی را متوجه روایی نمایند. تمرکز این قسمت بر تهدید ترکیبی انتخاب-بلوغ است، اما گهگاه گریزی به دیگر انواع تهدیدها نیز خواهیم زد.

نوع اول نتایج: هر دو گروه در یک جهت به طور تدریجی از یکدیگر فاصله می‌گیرند. یکی از الگوهای معمول انتخاب-بلوغ زمانی رخ می‌دهد که غیرهم‌ارزی اولیه میان گروهها با نرخهای میانگین متفاوتی، و در یک جهت رشد کرده، و از یکدیگر فاصله می‌گیرند (شکل ۵.۱). این الگو مدل بلوغ انتشار پنکه‌ای[۳۰۸] نام گرفته است؛ زیرا گروهها در طول زمان مانند پره‌های پنکه (از مرکز به محیط پنکه) از یکدیگر فاصله می‌گیرند. استاندارد کردن نمرات می‌تواند اثر انتشار پنکه‌ای را از میان ببرد چون انتشار پنکه‌ای محصول واریانسهای محاسبه‌شده‌ای است که به طور سیستماتیک در طول زمان افزایش می‌یابند. در جریان استاندارد شدن، نمرات به واریانس مربوطه‌شان تقسیم می‌شوند، و بنابراین نمرات در هر نقطه از زمان، در همان مقیاس قرار می‌گیرند و نه در مقیاسهای متفاوت. این الگو هماهنگ با اثرات مداخله است، با این وجود آیا می‌توان تفسیرهای جایگزین را تشخیص داده، و اثر آنها را از مطالعه خارج کرد؟



نمودار ۵.۱. اولین نتایج طرح گروه کنترل بدون‌مداخله، با پیش‌آزمون و پس‌آزمون

رالستون و همکارانش (Ralston et al., 1985) در مطالعه‌ای اثر ساعات کاری منعطف بر عمکرد کارکنان را در دو آژانس دولتی ایالتی مورد بررسی قرار دادند. در سازمانی که ساعات کاری منعطف نداشت، عمکلرد کارکنان در ابتدا پایین‌تر بود، و با گذشت زمان اندکی افزایش یافت. در سازمان با ساعات کاری منعطف، عملکرد در ابتدا بالاتر بود، و در طول زمان، با نرخ بیشتری افزایش یافت. این الگویی رایج در شبه‌آزمایشهاست، علی‌الخصوص زمانی که پاسخ‌دهندگان با انتخاب خودشان به گروههای آزمایش یا کنترل تخصیص داده می‌شوند. اما حتی

308 Fan-spread model of maturation

زمانی که پژوهشگر خود افراد را به گروهها تخصیص می‌دهد نیز، مداخله‌ها اغلب در اختیار افراد مستعد قرار می‌گیرد؛ افرادی که مشتاق رشد و بهتر شدن هستند و یا قابلیت بیشتری برای برقراری ارتباط با دیگران دارند. نکته آن است که اینگونه افراد به احتمال قوی به دلایلی که الزاماً ربطی به مداخله پژوهش ندارد، با نرخی سریعتر از دیگران رشد می‌کنند.

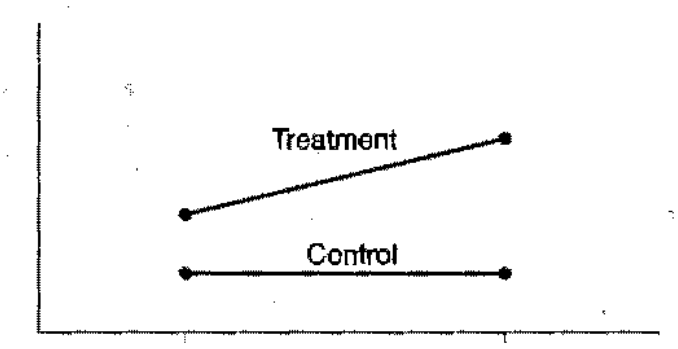
نشانه‌های محاسباتی متعددی وجود دارد دال بر این که گروههای غیرهم‌ارز با نرخهایی متفاوت در حال بالغ شدن هستند. اگر تفاوتهای میانگین گروهها نتیجه تهدید انتخاب-بلوغ باشد، رشد تفاضلی *میان* گروهها باید در *درون* گروهها نیز اتفاق بیافتد. این را می‌توان با استفاده از یک تحلیل درون‌گروهی که در آن، اعضای با عملکرد بالاتر با نمرات میانگین بالاتر باید سریعتر از اعضای با عملکرد پایینتر در همان گروه رشد کنند، شناسایی کرد. تهدید انتخاب-بلوغ اغلب با واریانسهای درون‌گروهی پس‌آزمونی‌ای همراه است که بزرگتر از واریانسهای معادل در پیش‌آزمون هستند. استفاده از نمودار پراکندگی نیز می‌تواند مفید باشد؛ پراکندگی نمرات پیش‌آزمون را نسبت به یک متغیر فرضی بلوغی (مانند سن یا سنوات تجربه) در نموداری ترسیم کنید. اینکار را برای گروه آزمون و گروه کنترل به طور مجزا انجام دهید. اگر خطوط رگرسیون حاصله تفاوت داشته باشد، احتمال وجود نرخ رشدهای مختلف وجود دارد. این تفاوتهای گروهی در شیب خط نمی‌تواند به دلیل مداخله باشد، چون تنها نمرات پیش‌آزمون مورد بررسی قرار گرفته است.

هیچ چیزی نمی‌تواند باعث شود تفاوت اولیه گروهی به طور خطی افزایش پیدا کنند، رشد در یک گروه ممکن است خطی باشد اما در گروه دیگر مربعی[۳۰۹]. اگرچه بر اساس تجربیات نگارندگان، بلوغ متفاوت از نوع انتشار پنکه‌ای متداول است. برای مثال، در زمینه آموزش، دانش‌آموزانی که پیشرفت تحصیلی بالاتری داشته‌اند، به صورت مداوم در مقیاسهای آزمون نمرات بهتری - نسبت به دیگر دانش‌آموزان که دارای پیشرفت تحصیلی پایین‌تر بوده‌اند- کسب می‌کنند. به باور نگارندگان، بلوغهای متفاوت با مدل انتشار پنکه‌ای در دیگر داده‌های طولی نیز وجود دارد. با این وجود، برخی فرمولسازی‌های نظری الگوهای انتخاب-بلوغ متفاوتی را پیش‌بینی می‌کنند، حتی در زمینه‌هایی مانند آموزش. برای مثال، بر اساس نظریه پیاژه، ناپیوستگی‌های تند و شیب‌داری می‌تواند در تفاوتهای رشد وجود داشته باشد؛ مانند زمانی که کودکی به طور ناگهانی مفهومی را درک می‌کند، و دیگری از انجام آن باز می‌ماند. بنابراین، هر مطالعه‌ای که طرح پایه به کار می‌گیرد، باید پیشفرضها و تفاوتهای بلوغی مختص به خود را [بسته به مقتضیات زمینه‌ای و نظری مطالعه] تعریف کند. بعضی اوقات، داده‌های پیش‌آزمون کمک ارزشمندی به تعریف این پیشفرضها می‌نماید. بعضی اوقات داده‌های بدست‌آمده از نمونه‌های مطالعات طولی نیز همین کارکرد را دارند، مانند ادعای این کتاب مبنی بر اینکه مدل انتشار پنکه ای غالباً در

309 Quadratic

داده‌های طولی مرتبط با دستاوردهای تحصیلی دیده می‌شود. اما همچنان، زمانهایی وجود دارند که فرضیه‌پردازی‌های نظری تنها منبع طراحی این پیش‌فرضهاست.

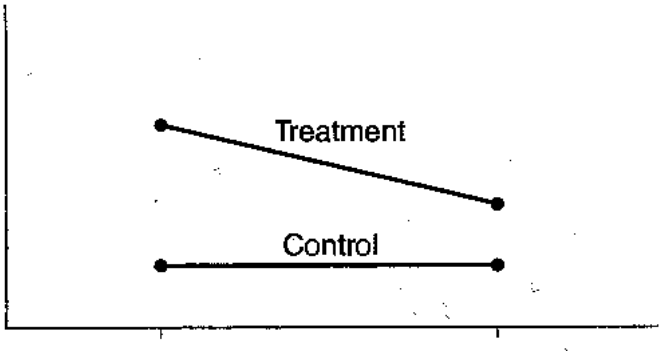
نوع دوم نتایج: هیچ تغییری در گروه کنترل مشاهده نمی شود. نارایانن و ناث (Narayanan & Nath, 1982) از این طرح برای بررسی اینکه چطور زمان کاری منعطف می‌تواند بر یک گروه از کارمندان أثر بگذارد استفاده، و این گروه را با دیگر همکارانشان در همان سازمان مقایسه کردند. نتایج نشان‌دهنده بهبود روابط مدیران و زیردستان در گروه با زمان کاری منعطف بود. بهبودی که در گروه کنترل مشاهده نشد (شکل ۵.۲).



نمودار ۵.۲. دومین نتایج از طرح گروه کنترل بدون‌مداخله با پیش‌آزمون و پس‌آزمون

هنگامی که تغییری در گروه کنترل دیده نمی‌شود، باید توضیحی وجود داشته باشد، که چرا رشد همزمان در گروه کنترل اتفاق نیافتاده است. غالباً توضیح اینکه چرا دو گروه (آزمون و کنترل) با نرخ متفاوتی رشد می‌کنند، و یا اینکه چرا هیچکدام از گروهها تغییری نکرده‌اند، ساده‌تر از آن است که بخواهیم علّت عدم‌تغییر گروه کنترل را با وجود تغییر گروه آزمون توضیح دهیم. برخی مواقع تحلیل‌های درون‌گروهی می‌تواند روشن‌کننده تهدیدهای میان‌گروهی باشد. مثلاً گروه آزمون به این علّت سریعتر دچار بلوغ می‌شود که سن گروه آزمون بالاتر از گروه کنترل است، در آن صورت، باید داده‌ها را بر اساس سن تقسیم کرد. اگر شرکت‌کنندگان گروه آزمون فارغ از سنشان، همچنان دچار بلوغ بیشتر می‌شوند، به این معناست که تهدید انتخاب-بلوغ در مورد مطالعه موردنظر چندان موضوعیت ندارد (این تهدید دلالت بر این دارد که رشد باید در یک گروه -و نه در دیگری- اتفاق بیافتد). با وجود تمام اینها، باید گفت که نمی‌توان اتکاء چندانی به این الگوی تفاوتهای تغییرات کرد. زیرا اینکه یک گروه بهبودیافته و دیگری تغییر نکند، الگوی ناشناخته‌ای نیست. بعلاوه، الگوی تفاوتهای تغییراتی که در اینجا به عنوان الگوی مهمتر مورد بحث قرار گرفت، تنها در کل چنین است، و با توجه به اینکه هر مطالعه تا حد زیادی وابسته به زمینه است، بسیاری از موارد معمول ممکن است برای یک مطالعه کاربرد نداشته باشند.

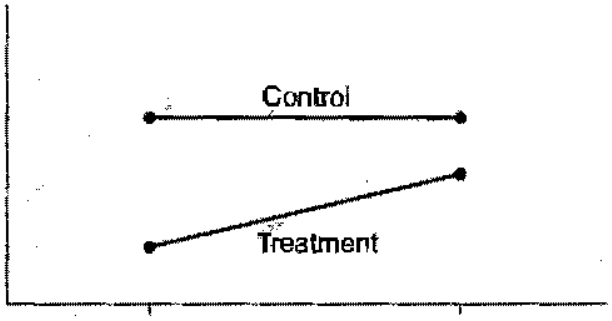
نوع سوم نتایج: برتری گروه آزمون در پیش‌آزمون، که با گذشت زمان از میان می‌رود. شکل ۵.۳ نشان‌دهنده حالتی است که برتری گروه آزمون در پیش‌آزمون در پس‌آزمون کاهش‌یافته، و یا ناپدید می‌شود. این نتایج، در مطالعه‌ای که اثر یکپارچه‌سازی مدارس را بر خودپنداره آکادمیک بررسی می‌کرد، رخ داد (Weber, Cook, & Campbell, 1971). این مطالعه نمونه‌ای متشکل از دانش‌آموزان سوم، چهارم و پنجم سیاهپوست داشت. در پیش‌آزمون، دانش‌آموزان سیاهپوست مدارسی که در آن تمامی دانش‌آموزان سیاهپوست بودند، خودپنداره آکادمیک بالاتری نسبت به دانش‌آموزانی داشتند که در مدارس یکپارچه (دارای دانش‌آموزان سفیدپوست و رنگین‌پوست) درس می‌خواندند. پس از آنکه فرایند یکپارچه‌سازی مدارس اتفاق افتاد، تفاوتهای اولیه از میان رفت.



نمودار ۵.۳. نتایج سوم از طرح گروه کنترل بدون‌مداخله با پیش‌آزمون و پس‌آزمون

برخی از تهدیدات روایی درونی مورد بحث در مورد شکلهای ۵.۲ و ۵.۱ در مورد شکل ۵.۳ هم مصداق دارد. اگرچه تهدید انتخاب-بلوغ در اینجا کمتر موضوعیت دارد، چون به ندرت اتفاق می‌افتد که کسانی که در ابتدا بسیار جلوتر از دیگران بوده اند، در مراحل بعدی عقب بمانند، و یا آنهایی که در ابتدا عقب‌تر از بقیه بوده اند، در ادامه به دیگران برسند. اگرچه که این می‌تواند اتفاق بیافتد. به عنوان نمونه، در یک زمینه آموزشی، اگر یک گروه از گروه دیگر از نظر سنی بزرگتر بوده اما از نظر هوشی پایین‌تر باشند، گروه مسن‌تر ممکن است بواسطه رجحان سنی، در مرحله اولیه عملکرد بهتری نسبت به گروه باهوشتر داشته باشد، اما رفته‌رفته این برتری با عملکرد گروه باهوشتر در مراحل بعدی ناپدید شود. اما چنین پدیده‌هایی نادر بوده، و در مثال وبر و همکارانش (Weber et al., 1971)، دو گروه از نظر سنی مشابه بودند. بنابراین بحث آن است که هیچکدام از فرایندهای بلوغ شناخته‌شده تا این زمان را نمی‌توان مسئول نتایج مشاهده شده در شکل ۵.۳ دانست. اگرچه ممکن است چنین فرایندهایی در آینده کشف و پیشنهاد شوند.

نوع چهارم نتایج: برتری گروه کنترل در پیش‌آزمون، که با گذشت زمان از میان می‌رود. در این مورد، مانند شکل ۵.۳، تفاوت میان آزمون و کنترل در پیش‌آزمون بزرگتر از پس‌آزمون است، اما در این حالت گروه آزمون در ابتدا ضعیفتر از گروه آزمون عمل می‌کند (شکل ۵.۴). این نتایج زمانی مطلوب است که مدارس مطالب مکمل را برای ارتقاء عملکرد گروههای ضعیف‌تر و محروم‌تر ارائه می‌کنند؛ و یا زمانی که سازمانی تلاش می‌کند تا برای ارتقاء عملکرد واحد ضعیف‌تر تغییراتی اعمال کند. کلر و هولاند (Keller & Holland, 1981) در حین تلاش برای ارزیابی اثر تغییرات شغلی بر عملکرد شغلی، نوآوری، رضایت و یکپارچگی، در سه سازمان پژوهش و توسعه، این الگو را پیدا کردند. کارکنانی که بهبود شغلی می‌گرفتند یا به شغل دیگری انتقال داده می‌شدند، گروه آزمون را تشکیل می‌دادند، و دیگر کارکنان سازمان گروه کنترل در نظر گرفته می‌شدند. نتایج با فاصله یکسال دو بار محاسبه می‌شد. اگرچه، این پژوهش هیچ تمرکز ترمیمی یا اصلاحی [برای واحدی ضعیف یا محروم] نداشت، داده‌ها الگوی موردبحث (شکل ۵.۴) را نشان می‌داد، و نتایج آنهایی که تغییرات شغلی را تجربه کرده بودند بهبود یافته، در حالی که نتایج دیگران بدون تغییر مانده بود.



نمودار ۵.۴. چهارمین نتایج از طرح گروه کنترل بدون‌مداخله با پیش‌آزمون و پس‌آزمون

در این حالت، نتایج مبتلا به تهدیدات نوعی مرتبط با مقیاس‌سازی (مانند انتخاب-ابزار) و گذشت زمان محلی (انتخاب- گذشت زمان) خواهد بود. اما دو عنصر خاص بارزتر است. اول، اگر شرکت شغل آن دسته از کارکنان را که در پیش‌آزمون عملکرد ضعیفی داشتند، تغییر داده باشد، نتایج آن کارکنان باید در پس‌آزمون بهتر شود (رو به بالا حرکت کند[۳۱۰])؛ نتیجه‌ای که نموداری شبیه شکل ۵.۴ را تولید می‌کند. اگر تفاوت گروههای آزمون و کنترل در مطالعه کلر و هلاند (۱۹۸۱) در طول زمان پایدار می‌بود (چیزی که با طرح فعلی نمی‌توانیم در مورد آن قضاوتی داشته باشیم، اما اگر دو پیش‌آزمون می‌داشتیم امکان‌پذیر بود)، در آنصورت رگرسیون نمی‌توانست تهدیدی به حساب بیاید. بنابراین، در طرحهای گروه کنترل غیرهم‌ارز لازم است تا بدانیم علّت تفاوت اولیه

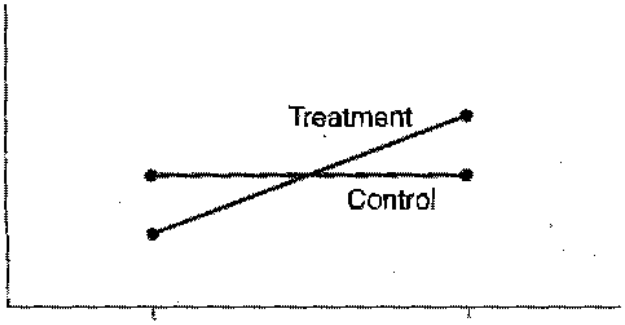
310 Regressed upward

گروهها چه بوده است. این شامل دانستن چرایی تمایل برخی افراد برای تخصیص خود به یک مداخله، و نه مداخله دیگر نیز می‌شود.

دومین عنصر خاص این طرح آزمایشی آن است که نتایج نشان داده شده در شکل ۵.۴، مدل انتخاب-بلوغ نوع انتشار پنکه‌ای را منتفی ساخته، و یا نشان می‌دهد که اگر چنین اثری وجود داشته باشد، مداخله قادر است بر آن فائق بیاید. با این وجود، دیگر انواع الگوهای انتخاب-بلوغ مجال بروز دارند. برای مثال در مطالعه کلر و هولاند (۱۹۸۱)، کسانی که شغل آنها تغییر کرد ممکن است جزء کارکنان تازه‌کار سازمان بوده باشند، که این می‌توانسته دلیل نمرات پایین‌تر آنها در پیش‌آزمون باشد. اما در عین حال، آنها ممکن است بیشتر مستعد یادگیری از تجربیات باشند، چیزی که باعث می‌شود عملکرد آنها بی‌تناسب با گروه دیگر، سریعتر بهبود یابد. برای در نظر گرفتن این احتمال لازم است تا داده‌های مرتبط با سن و سابقه کار در سازمان شرکت‌کنندگان مورد تحلیل قرار بگیرد. به طور کلی، این نتایج را می‌توان به طور علّی تفسیر کرد. اما در صورتی که مشخصات خاص یک مطالعه، موجب شکلگیری الگوی انتخاب-بلوغ پیچیده‌ای شود، این مسأله باید در هر مطالعه به طور دقیق مورد بررسی قرار گیرد.

نوع پنجم نتایج: نتایجی که در جهتی همراستا با روابط از روی یکدیگر می‌گذرند و یکدیگر را قطع می‌کنند. در نتایج فرضی نمایش داده شده در شکل ۵.۵، خطوط روندها یکدیگر را قطع می‌کنند، و میانگینها به طور پایداری در پیش‌آزمون در یک جهت متفاوت هستند و در پس‌آزمون در دیگر جهت مخالف. این نتایج می‌توانند به طور ویژه‌ای نشان دهنده رابطه علّی باشد. اولاً، موجّه‌بودن تهدید انتخاب-ابزار کاهش پیدا می‌کند، چون انجام هیچ تغییر شکلی[۳۱۱] در داده‌ها نمی‌تواند برهم‌کنش (یا تقاطع) را از میان ببرد. به عنوان مثال، اثر سقف نمی‌تواند توضیح دهد که چرا گروهی که در ابتدا نمرات پایین‌تری داشت، از گروهی در ابتدا نمراتی بالاتر داشت عبور کرده، و پیشی گرفته است. یک مصنوع مقیاس‌سازی قانع‌کننده باید چنین فرض کند که نمره میانگین پس‌آزمون گروه مداخله به این دلیل افزایش یافته که مشخصات فاصله‌ای آزمون، تغییرات را در نقاطی از مقیاس که فاصله بیشتری از میانگین دارد، آسانتر می‌سازد. گرچه این حادتر شدن یک اثر واقعی را توضیح می‌دهد، و نه خلق یک اثر کاملاً مصنوعی را.

311 Transformation



نمودار ۵.۵. پنجمین نتایج از طرح گروه کنترل بدون‌مداخله با پیش‌آزمون و پس‌آزمون

دوم، در نمودارهایی مانند نمودار ۵.۵. تهدیدهای انتخاب-بلوغ کمتر احتمال بروز دارند، زیرا الگوهای بلوغ برهم‌کنشی متقاطع، اگرچه که اتفاق می‌افتند، اما چندان موردانتظار نیستند. نمونه‌ای از الگوی نمایش داده شده در نمودار ۵.۵ در مطالعه کوک و همکارانش (Cook et al., 1975) در مورد بازتحلیل داده‌های بدست آمده در زمینه اثربخشی طرح برنامه خیابان کنجد مشاهده شد. آنها دریافتند که دانش‌آموزانی که تشویق به تماشای برنامه شده بودند، در پیش‌آزمون به طور پایداری کمتر از دانش‌آموزانی که به تماشای برنامه تشویق نشده بودند اطلاعات داشتند، اما دانش آنها در پس‌آزمون به طور پایداری بالاتر از گروه کنترل بود. اما آیا دانش‌آموزان گروه آزمون نسبت به افراد گروه کنترل جوانتر و باهوشتر بودند و در نتیجه، در پیش‌آزمون نمرات پایینتری کسب کرده، و به مرور زمان به دلیل توانایی‌های بیشتر خود سریعتر تغییر کرده بودند؟ خوشبختانه داده‌ها نشان داد که دو گروه آزمون و کنترل از نظر سن و یا چندین مقیاس قابلیت و توانایی‌های اندازه‌گیری شده، در پیش‌آزمون تفاوت چندانی نداشتند، که این خود از موجّه‌بودن این تهدید می‌کاست.

سوم، نتایج شکل دهنده نمودار ۵.۵ احتمال بروز تهدید رگرسیون را منتفی می‌سازند. گرین و پادساکف (Green & Podsakoff, 1978) هنگامی که نحوه اثر حذف برنامه انگیزه مالی را بر رضایت کارکنان در کارخانه تولید کاغذ مورد بررسی قرار می‌دادند، چنین نمودار متقاطعی از داده‌ها را مشاهده کردند. کارکنان به سه دسته با عملکرد بالا، متوسط و پایین تقسیم‌بندی شدند، و میزان رضایتمندی قبل و بعد از انگیزش مالی مورد اندازه‌گیری قرار گرفت. پس از حذف انگیزه‌های مالی، رضایتمندی گروه با عملکرد بالا به طور پایداری کاهش، و رضایتمندی گروه با عملکرد پایین افزایش یافت، و رضایتمندی گروه متوسط تغییری پیدا نکرد. این تفاوت در شیبها می‌توانست به دلیل رگرسیون باشد، اگر هر سه گروه در میانگین مشترکی تلاقی کنند (چیزی شبیه نمودار ۵.۴). اما رگرسیون آماری نمی‌تواند توضیح دهد چرا افراد با عملکرد پایین در پس‌آزمون به طور پایداری افراد با عمکلرد بالا را پشت‌سرگذاشتند (اگرچه که رگرسیون می‌توانسته تخمینهای مداخله را افزایش داده باشد).

متأسفانه، هر گونه تلاش برای داشتن طرحی که تولیدکننده نتایجی شبیه نمودار ۵.۵ باشد، واجد ریسکی قابل‌توجه است. یک دلیل آن است که توان شناسایی یک برهم‌کنش[۳۱۲] آماری پایدار اندک است (Aiken & West, 1991). در نتیجه، اینگونه مطالعات باید با دقت فراوانی طراحی شوند. این خصوصاً زمانی مصداق دارد که یک فرایند انتشار پنکه‌ای مانند آنچه در شکل ۵.۲ نشان داده شد، موردانتظار باشد، زیرا در آن زمان است که یافته‌ی عدم‌وجود تفاوت باعث می‌شود بتوان فهمید که آیا مداخله بی‌اثر بوده است، و یا اینکه دو نیروی متعارض (نیروی مداخله و نیروی انتشار پنکه‌ای) یکدیگر را خنثی کرده‌اند. حتی اگر تفاوتی در شیب خطها وجود می‌داشت، نمودار می‌بایست چیزی شبیه شکل ۵.۴ می‌بود و نه شکل ۵.۵؛ و در نتیجه، شکل ۵.۴ غیرقابل تفسیر خواهد بود. بنابراین، محققین نباید برای بدست آوردن نتایجی مشابه نمودار ۵.۵، بر طراحی اتکا کنند. بلکه در عوض، باید اقداماتی در راستای اضافه کردن کنترلهای طراحی قویتر به طرح پایه پیش‌آزمون-پس‌آزمون با گروه کنترل اتخاذ شود.

راههایی برای بهبود طرح گروه کنترل دستکاری‌نشده با نمونه‌های وابسته پیش‌آزمون و پس‌آزمون

مانند طرح‌های ارائه شده در فصل قبل، این طرح پایه را نیز می‌توان با اضافه کردن عناصر طراحی خاصی که برای از بین بردن تهدیدهای روایی موجّه در آن زمینه آزمایش کارا هستند، تا حد قابل‌توجهی ارتقاء داد. از جمله این موارد می‌توان به موارد زیر اشاره کرد.

استفاده از دو پیش‌آزمون. در این حالت همان پیش‌آزمون را دو بار، و در دو زمان متفاوت تکرار می‌کنیم؛ ترجیحاً با همان فاصله زمانی‌ای که پیش‌آزمون دوم نسبت به پس‌آزمون دارد. این طرح را به شکل زیر نشان می‌دهند.

NR	O_1	O_2	X	O_3
NR	O_1	O_2		O_3

انجام دو پیش‌آزمون به محقق اجازه می‌دهد تا سوگیری‌های احتمالی در تحلیل مداخله اصلی را بهتر درک نماید. اگر در تحلیل فاصله زمانی میان نقطه زمانی O_1 به O_2 ، «اثر مداخله» مشاهده شود، سوگیریهای مشابهی در فاصله زمانی میان O_2 به O_3 نیز می‌تواند وجود داشته باشد. ورتمن و همکارانش (Wortman et al., 1978) طراحی مشابهی را برای آزمایش اثر کوپن آموزش راک‌الوم بر نمرات آزمون روخوانی بکار بردند. در این برنامه، والدین یک مدرسه محلی را برای فرزند خود انتخاب می‌کردند، و کوپنی معادل هزینه آموزش در آن مدرسه دریافت می‌کردند. هدف از این برنامه، افزایش رقابت میان مدارس درون این سیستم بود. تحلیل اولیه داده‌ها

312 Interaction

توسط دیگران نشان می‌داد که دادن کوپنها عملکرد آکادمیک را کاهش می‌دهد. اما ورتمن و همکارانش در مورد نتایج بدست‌آمده دچار تردید بودند. در نتیجه، یکی از گروههای دانش‌آموزان را در طول نمرات مرحله اول تا سوم در مدارس کوپنی و غیرکوپنی دنبال کرده، و نمرات را با استفاده از یک پیش‌آزمون دوگانه بازتحلیل کردند. بعلاوه، آنها مدارس کوپنی را به دو دسته مدارس واجد و فاقد سیستم کوپنی سنتی تقسیم کردند. پیش‌آزمون اضافی به محققین اجازه می‌داد رشد نمرات خواندن قبل از مداخله (در فاصله زمانی میان O_1 به O_2) را با تغییر در نرخهای پس‌آزمون (در فاصله زمانی میان O_2 به O_3) مقایسه کنند. و به همین دلیل، کاهش در نمرات خواندن که پیش از این به مدارس کوپنی نسبت داده شده بود، تنها به گروه کوپن غیرسنتی نسبت داده شد. گروههای دریافت‌کننده کوپن سنتی و غیر کوپنی اثر متفاوتی نشان ندادند- اثر متفاوتی که نتوان آنرا با استفاده از امتداد همان نرخهای بلوغی که قبلاً در مدارس سنتی و مدارس کنترل کوپنی دیده شده بود، توضیح داد. داشتن دو پیش‌آزمون ارزیابی تهدید انتخاب-بلوغ را امکان‌پذیر می‌کند، با این فرض که نرخهای میان O_1 به O_2 در فاصله زمانی میان O_2 به O_3 نیز ادامه خواهد یافت. پیش‌فرضها تنها برای گروه دستکاری‌نشده قابل بررسی است. بعلاوه، با در نظر گرفتن خطای محاسبه، نرخ‌های رشد میان‌گروهی بادقت اندکی تخمین‌زده خواهد شد، و تغییرات ابزار می‌تواند باعث شود رشد محاسبه‌شده میان O_1 به O_2 مشابه O_2 به O_3 نباشد. در نتیجه، داشتن دو پیش‌آزمون با گروههای غیرهم‌ارز، کامل و بی‌نقص نیست. بلکه، دومین پیش‌آزمون تا حد زیادی در ارزیابی احتمال موجّه‌بودن تهدید انتخاب-بلوغ موثر باشد، چون می‌تواند نشان‌دهنده تفاوتهای رشد پیش از مداخله باشد. وجود دو پیش‌آزمون همچنین می‌تواند کمک کند تا اثر رگرسیون را شناسایی کنیم، اگر مشاهدات انجام شده در O_2 در هر یک از گروهها، در مقایسه با O_1 به طور غیرمعمولی بالا یا پایین باشد. همچنین کمک می‌کند تا همبستگی میان مشاهدات انجام شده در زمانهای مختلف را به طور دقیقتری تخمین بزنیم؛ چیزی که در تحلیل آماری بسیار ارزشمند است. بدون نقاط زمانی اندازه‌گیری، همبستگی میان O_2 به O_3 در گروه دستکاری‌شده، تخمینی مبهم از همبستگی‌ای خواهد بود که می‌توانست در غیاب مداخله وجود داشته باشد. پس چرا پیش‌آزمونهای متعدد بیشتر مورد استفاده قرار نمی‌گیرند؟ ناآگاهی می‌تواند یک دلیل باشد، اما دلیل دیگر آن است که برخی اوقات توجیه اقتصادی ندارد. اغلب اوقات اگر بخت با محقق یار باشد، تنها قادر است مداخله را به اندازه انجام یک پیش‌آزمون به تعویق بیاندازند، چه رسد به انجام دو پیش‌آزمون، و یا انداختن فاصله زمانی میان پیش‌آزمون و پس‌آزمون. برخی مواقع آرشیوها امکان انجام پیش‌آزمون دوم و بیشتر و در نتیجه انجام طرحهای سری‌زمانی قویتری را فراهم می‌آورند. به علاوه، برخی اوقات افراد صادرکننده مجوز هزینه‌کردهای تحقیق، دوست ندارد ببینند پولها برای چیزی غیر از اندازه‌گیریهای پس‌آزمون خرج شود. معمولا مجاب کردن این افراد درباره مفید بودن وجود پیش‌آزمونها و گروههای کنترل کار دشواری است. حال خودتان

دشواری داشتن پیش آزمون اضافی را در نظر بگیرید. با این وجود، هر زمان که سیستم آرشیوی، چارچوب زمانی، منابع و سیاستهای اجازه می‌دهند، باید دو پیش‌آزمون پیش از مداخله اجرا کنید.

استفاده از تکرارهای جابجا شونده. با تکرارهای جابجاشونده محقق مداخله را در تاریخی (زمانی) دیرتر، بر روی گروهی که در ابتدای مطالعه به عنوان گروه کنترل مورد استفاده قرار گرفته بودند، اجرا می‌کند. این طرح را به صورت زیر نشان می‌دهند:

NR	O_1	X	O_2		O_3	
NR	O_1		O_2	X	O_3	

بسادور و همکارانش (Besadur et al., 1986) شکلی از این مطالعه را برای بررسی نحوه اثر آموزش ضمن خدمت بر نگرش مهندسان نسبت به تفکر سنت‌شکنانه در هنگام حل مسأله، مورد استفاده قرار دادند. اندازه‌گیری یکبار قبل از انجام مداخله، یکبار بعد از آموزش گروه اول از مهندسین، و نهایتاً بعد از آموزش گروه غیرهم ارز دوم انجام شد. گروه دوم در فاز اول، نقش گروه کنترل را ایفاء می‌کردند، در حالی که در فاز دوم نقشها جابجا شدند. اگرچه، فاز دوم عیناً تکرار فاز اول نیست. شرایط زمینه‌ای پیرامون مداخله دوم متفاوت از مداخله اول است، هم با توجه به گذشت زمان، و هم چون اثر مداخله از گروه اول حذف شده است. حتی اگر اثر حذف نشده باشد، فرض بر این است که مداخله مربوطه در حال حاضر اثری ندارد (این طرح حتی زمانی که اثر مداخله اول استمرار داشته باشد نیز همچنان قابل استفاده است، خصوصاً اگر گروه دوم در هنگام دریافت مداخله از نظر عمکلرد به گروه آزمون رسیده باشد). با توجه به تفاوت زمینه‌ای میان مداخله اول و دوم، انجام مداخله دوم یک تکرار[۳۱۳] تغییریافته است، که روایی درونی و روایی بیرونی اینکه آیا زمینه جدید اثر مداخله را تغییر می‌دهد یا نه، را بررسی می‌کند.

این طرح را می‌توان برای بیش از دو گروه نیز انجام داد. در این حالت، می‌توان گروهها را به صورت تصادفی به زمانهای مختلفی که مداخله را شروع می‌کنند تخصیص داد، زیرا اگر قرار است طرح با چندین گروه انجام شود، بر اساس تعریف، باید زمانهای تصادفی پشت‌سرهم و بدون وقفه متعددی وجود داشته باشد. عنصر تصادفی بودن در این طرح می‌تواند استنباطها را تقویت نماید. این امر در زمانی که گروههای بیشتر، و زمانهای آزمون بیشتری وجود دارد، اهمیت بیشتری پیدا می‌کند. اما حتی بدون تخصیص تصادفی مداخلهها به فاصله‌های زمانی نیز، هنگامی که گروهها و نقاط زمانی بیشتری در طرح وجود دارد، فرصتهای تحلیل به طور کارایی افزایش می‌یابند (مانند Koehler & Levin, 1998).

313 Replication

مهمترین محدودیتهای این طرح از این واقعیت ناشی می‌شوند که موارد بعدی که به عنوان کنترل عمل می‌کنند، یا (۱) همان مداخله را حفظ می‌کنند، اما فرض را بر این می‌گذاریم که مداخله اثر ناپیوسته بلندمدتی هم‌جهت با مداخله‌ای که بعداً روی گروههای کنترل اولیه اجرا شده ندارد، و یا اینکه (۲) مداخله را از گروه آزمون اولیه حذف می‌کنیم. این بطور بالقوه، فرایندهای رقابت جبرانی و موارد مانند آن را موجب می‌شود، فرایندهایی که باید به طور دقیق توضیح داده شده، اندازه‌گیری شود، و در تحلیلها لحاظ شود. گذشته از اینها، تکرارهای جابجا شونده طرحی قدرتمند است. تنها الگوهایی از تغییرات رخ داده در طول زمان، که شبیه ترتیب زمانی اجرای مداخله هستند، می‌توانند به عنوان یک تفسیر جایگزین به کار گرفته شوند.

بکارگیری گروه کنترلی با مداخله معکوس. این طرح به صورت زیر نمایش داده می شود:

NR	O_1	X_+	O_2
NR	O_1	X_-	O_2

X_+ نشان‌دهنده مداخله‌ای است که انتظار می‌رود اثری با جهتی معین داشته باشد، و X_- نشان‌دهنده مداخله‌ای معکوس است که انتظار می‌رود اثری عکس (مداخله اول) داشته باشد. هکمن و همکارانش (Hackman et al., 1978) این طرح را برای بررسی اینکه چطور تغییرات در محتوی انگیزشی شغل می‌تواند بر نگرش و رفتار کارگران اثر بگذارد، بکار گرفتند. در نتیجه، یک نوآوری تکنولوژیکی، شغلهای دفتری در بانک به گونه‌ای تغییر داده شدند که کار روی برخی واحدها پیچیده‌تر و چالش برانگیزتر (X_+)، و بالعکس، کار روی برخی دیگر از واحدها ساده‌تر (X_-) شدند. این تغییرات بدون آنکه به کارکنان در مورد نتایج انگیزشی چیزی گفته شود، انجام شد، و اندازه‌گیری مشخصات شغلی، نگرش کارکنان و رفتار کارکنان قبل و بعد از بازتعریف شغل انجام گرفت. اگر مداخله (X_+) نمرات گروه آزمون را ارتقا می‌داد، و مداخله (X_-) نمرات گروه مقابل را کاهش می‌داد، در نتیجه، باید یک برهم‌کنش آماری پدیدار شود، که حاکی از اثر مداخلهست.

طرح مداخله معکوس می‌تواند یک مزیت خاص مرتبط با روایی سازه داشته باشد. سازه علّی باید به طور مستحکمی تعیین و دستکاری شود تا بتوان آزمونی حساس داشت، که در آن، یک ویرایش از علّت (غنی کردن شغل) یک گروه را در یک جهت تغییر داده، در حالی که مخالف مفهومی آن (تهی کردن شغل)، گروه مقابل را در جهت عکس تغییر می‌دهد. برای فهم بهتر این موضوع، تصور کنید اگر هکمن و همکارانش تنها یک گروه شغلی غنی‌شده داشتند، و از گروه کنترل استفاده نمی‌کردند. شیب تندتر پیش‌آزمون-پس‌آزمون در شرایط شغل غنی‌شده می‌تواند به تغییرات شغلی، و یا احساس پاسخ‌دهندگانی که فرضیات را حدس زده بودند، نسبت داده

شود. اگر کاهش موردانتظار پیش‌آزمون-پس‌آزمون در رضایت شغلی در گروه معکوس مداخله نیز مشاهده شود، چنین گزینه‌هایی کمتر موجّه خواهند بود، زیرا آگاهی نسبت به بودن در پژوهش معمولاً باعث می‌شود پاسخ‌دهندگان جوابهایی مقبول به لحاظ اجتماعی ارائه کنند. برای آنکه بتوان هم افزایش در گروه با شغل غنی‌شده، و کاهش در گروه معکوس را توضیح داد، هر مجموعه از پاسخ‌دهندگان می‌بایست فرضیات را به شیوه‌ای متناسب با وضعیت خودشان حدس‌زده باشند.

تفسیر این طرح، به تولید دو اثر با جهت‌های (علامت) مخالف وابسته است. بنابراین چنین فرض می‌شود که تغییرات انگیزشی و زمانی به صورت دیگری می‌توانست رخ دهد. مواقعی که تغییرات در گروههای آزمون و کنترل متفاوت اما در یک جهت باشد، به سختی می‌توان نتایج را تفسیر کرد، زیرا رابطه آنها با یک گروه کنترل دستکاری نشده نامشخص است. اضافه کردن چنین گروه کنترلی مفید بوده، و در صورت اقتصادی بودن بهتر است انجام شود. همچنین در بسیاری از زمینه‌ها ملاحضات کاربردی و اخلاقی مانع از بکارگیری گروه کنترل معکوس می‌شود. اغلب مداخلهها اهداف اصلاحی و در راستای بهبود جامعه دارند، در نتیجه یک مداخله معکوس می‌تواند آسیب‌زا باشد. اگرچه این مسأله در مورد مطالعه هکمن و همکارانش چندان جدی نبود. چه کسی می‌تواند بگوید کدامیک از این استراتژیها -یعنی پیچیده‌تر کردن یا ساده‌تر کردن شغل یک فرد- می‌تواند سودمندتر باشد؟

اندازه‌گیری مستقیم تهدیدات روایی. این اندازه‌گیریها به محقق اجازه می‌دهد تا وجود احتمالی تهدیدات روایی را تشخیص دهد. در مطالعه نارایانان و ناث (Narayanan & Nath, 1982)، در یک شرکت برای گروهی از کارکنان زمان کاری انعطاف‌پذیر مقرر شد، و دیگر کارکنان به عنوان گروه کنترل به حساب می‌آمدند. در صورتی که سرپرستان رفتار متفاوتی نسبت به دو گروه داشته باشند، خطر گذشت زمان می‌تواند روایی را تهدید کند. برای بررسی احتمال وجود این خطر، این دو محقق وجود چنین تغییراتی را اندازه‌گیری کرده، و اثری از آنها نیافتند. البته این تنها مثالی از تهدید گذشت زمان است، و مثالهای فراوانی دیگری از این دست را می‌توان ذکر کرد. بنابراین محققین باید محتاط باشند تا منتفی شدن احتمال یک خطر، نتواند آنها را نسبت به موجّه بودن خطری دیگر بی‌توجه نماید. هر یک از خطرات باید مفهوم‌سازی شده، و با روایی اندازه‌گیری و تحلیل شوند. این باعث می‌شود اندازه‌گیری و محاسبه تهدیدات دشوار باشد. با این وجود، اندازه‌گیری تهدیدات می‌تواند تحلیلهای بعدی را تسهیل نماید، چون اجازه می‌دهد تفسیرهای جایگزین با تحلیلهای اولیه‌ای که برای درک ناهمگونی‌های میان گروههای آزمون و کنترل انجام می‌شود، ترکیب شوند.

جفت‌سازی از طریق گروههای کنترل همتایان[۳۱۴]

314 cohort

در بسیاری از مؤسسات به طور دائم چرخش‌هایی وجود دارد. مثلاً گروهی از دانشجویان که فارغ التحصیل می‌شوند و گروهی دیگر جای آنها را می‌گیرند. مدارس مثالی روشن در این رابطه هستند، زیرا کودکان از مقطعی به مقطع بالاتر ارتقاء پیدا می‌کنند. مثال دیگر، کسب و کارهایی هستند که در آنها گروهی از کارآموزان با گروهی دیگر جایگزین می‌شوند، خانواده‌هایی که یک بچه بعد از بچه‌ای دیگر می‌آید، و زندانهایی که یک گروه زندانی، به دنبال گروه دیگر وارد می‌شوند. کلمه همتایان به گروههای جانشینی که در چنین فرایندهایی حرکت می‌کنند، اطلاق می‌شود. استفاده از همتایان به عنوان گروه کنترل، خصوصاً زمانی می‌تواند مفید باشد که (۱) یک دوره تجربیاتی داشته‌اند که گروه قبل یا بعد از آنها نداشته‌اند، (۲) همتاها تنها از جهاتی جزئی با همتایان دوره‌های قبل و بعد خود تفاوت دارند، (۳) سازمانها اصرار داشته باشند که یک مداخله روی تمامی افراد اجرا شود، و در نتیجه، امکان داشتن همزمان گروههای کنترل را از بین می‌برند، و تنها امکان باقیمانده، انتخاب گروههای کنترل از همتایان قبلی یا بعدی خواهد بود، و در نهایت، (۴) رکوردهای آرشیوی یک سازمان را بتوان برای ساختن و سپس مقایسه گروهها مورداستفاده قرار داد.

پیشفرض مهم در مورد همتایان آن است که تفاوتهای مبنای انتخاب میان همتایان، کمتر از تفاوتهایی است که میان گروههای غیرهمتا مشاهده می‌شود. اگرچه باید وجود این پیشفرض در هر مطالعه مورد بررسی قرار بگیرد؛ برای مثال از طریق تحلیل مشخصاتی در زمینه مطالعه که احتمالاً با نتایج همبستگی دارند. اما حتی با انجام این تحلیلها نیز قیاس‌پذیری گروهها [آزمون و کنترل] هیچگاه به اندازه قیاس‌پذیری گروههای بدست‌آمده از تخصیص تصادفی نخواهد بود. علاوه بر این، در مطالعه‌ای مروری بر روی تحقیقات ژنتیک رفتاری انجام‌شده در حوزه عملکرد ذهنی، مشخص شد تفاوتهای محیطی در محیط خردی که خواهر و بردارها در اطراف خود دارند و یا ایجاد می‌کنند، باعث می‌شود که دو کودک انتخاب شده از یک خانواده به همان میزان متفاوت باشند که دو کودک جفت‌شده[۳۱۵] انتخاب شده در جریان تخصیص تصادفی (Plomin & Daniels, 1987). اگر این نتایج درست و قابل‌تعمیم به دیگر زمینه‌های غیرذهنی باشد، باعث می‌شود نتوان جایگاه خاصی برای فرزندان یک خانواده به عنوان همتایان قائل شد. اما با این وجود، طرحهای گروه کنترل از فرزندان یک خانواده، همچنان به عنوان یکی از تکنیکهای محبوب برای بسیاری از اقتصاددانان در هنگام مطالعه اثر متغیرهای بیرونی بر مشارکت یا پذیرش آموزشی نیروی کار است (Aronson, 1998; Ashenfelter & Krueger, 1994; Currie & Duncan, 1995, 1999; Duncane, Yeung, Brooks-Gunn, & Smith, 1998; Geronimus & Korenman, 1992).

به عنوان نمونه می‌توان به مطالعه مینتون (Minton, 1975) که در آن از گروه کنترل خواهر و برادران استفاده شده است، اشاره کرد. این مطالعه به بررسی این موضوع می‌پرداخت که فصل اول برنامه خیابان کنجد چه اثری بر نمرات آمادگی فرهنگ شهرنشینی نمونه‌ای ناهمگون از کودکان مهدکودکی داشته است. مینتون کودکستانی

315 Paired

که در آن، در انتهای اولین سال حضور کودکان، تست مذکور بر روی آنها صوت گرفته بود را انتخاب کرد. برای انتخاب یک گروه کنترل، نمرات تست خواهرها و برادران بزرگتر این کودکان را، که قبل از پخش برنامه خیابان کنجد به همین کودکستان می‌آمدند، را در نظر گرفتند. بنابراین محقق نمراتی را داشت که مربوط به زمانی بود که خواهران و برداران بزرگتر این کودکان در سنی مشابه سن آن کودکان در زمان پخش برنامه خیابان کنجد بودند. این طرح به صورت زیر نشان داده می‌شود:

NR O_1

NR X O_2

خط نقطه چین (.....) وسط نشان‌دهنده یک گروه کنترل همتای غیرهم‌ارز است. در اینجا، ما ابتدا یک طرح با گروه کنترل همتا و بدون پیش‌آزمون را معرفی کرده، و در قسمت بعدی یک پیش‌آزمون به آن می‌افزاییم.
نمرات اندیسها نشان‌دهنده زمان محاسبه‌های انجام‌شده هستند، و اثر با مقایسه O_1 و O_2 ارزیابی می‌شوند. این طرح به روشنی نشان می‌دهد که گروه خواهران و برادران بزرگتر به عنوان گروه کنترل هم‌سن (مشابه گروه آزمون)، با همان درجه بلوغ، و بی‌نیاز از انتخاب[۳۱۶] مورد استفاده قرار می‌گیرند.
علیرغم شباهتهای میان همتایان در سطح بلوغ، و دیگر متغیرهای مرتبط با رابطه خانوادگی، اکتفا کردن صرف به مقایسه این دو مشاهده (اندازه‌گیری)، آزمونی ضعیف برای فرضیات علّی ارائه خواهد کرد. اولاً، همچنان نوعی مشکل انتخاب باقی می‌ماند، زیرا خواهران و برادران بزرگتر از نظر سنی احتمالاً از نظر نمرات شناختی از عملکرد بهتری برخوردارند (Zajonc & Markys, 1975). یک راه برای کاهش این خطر آن است که داده‌ها را به طور مجزا، و بر اساس ترتیب تولّد کودکان بزرگتر تحلیل کنیم؛ زیرا اثر ترتیب سنی با افزایش ترتیب تولد برادران بزرگتر تعدیل می‌شود (Zajonc & Markus, 1975). این طرح نسبت به اثر گذشت زمان نیز آسیب‌پذیر است؛ کودکان بزرگتر و کوچکتر ممکن است رویدادهای متفاوتی به غیر از تماشا یا عدم‌تماشای برنامه خیابان کنجد را تجربه کرده باشند. این رویدادهای متفاوت تجربه شده می‌تواند روی سطح دانش آنها تاثیرگذار باشد.
یک راه‌حل برای تشخیص این مشکل آن است که همتایان را به دسته‌هایی تقسیم کنیم که با فاصله ۱، ۲، ۳ و یا بیشتر سال، قبل از همتاییان خود کودکستان را تجربه کرده‌اند. از این طریق، می‌توان بررسی کرد که آیا یادگیری بیشتر گروه جوانتر، در جریان تمام رویدادهای متفاوت تاریخی که این گروه تجربه کرده است، همچنان پایدار بوده؟ حتی در این صورت نیز این رویه نخواهد توانست رویدادهایی را که در طول همان سال پخش برنامه خیابان کنجد رخ داده، را کنترل کند. بنابراین، راه‌حل بهتر آن است که آزمایش را در مدارس مختلف و در

316 Selection-reduced

سالهای مختلف تکرار کنیم. اگر اثر هر بار دیده شود، به این معنی است که هر رویداد رخداده در گذشته یا رویدادهایی که دارای اثری بوده که با اثر مداخله اشتباه گرفته شده، می‌بایست به طور موقت از مدرسه‌ای به مدرسه‌ی دیگر در طول این سالها از مداخله تقلید کرده باشد. البته در مورد این برنامه، با توجه به محبوبیت اولیه برنامه در خانه‌های خصوصی، حتی این امکان آخر نیز شدنی[۳۱۷] نبود. بنابراین هیچ گروه از مدارس با حداقل سطح درمعرض اثر بودن، امکان‌پذیر نبود.

برخی اوقات مدیریت مستقیم می‌تواند به ارزیابی تهدیدهای گذشت زمان و انتخاب کمک نماید. برای مثال، دوین و همکارانش (Devine et al., 1990) شبه‌آزمایشی را برای بررسی اثر کارگاه آموزشی مراقبت رواندرمانی بر توجه و مراقبت پرستاران نسبت به بیمارانی که جراحی کیسه صفرا انجام داده‌اند، و متعاقباً بهبودی این بیماران، انجام دادند. برای هفت ماه پیش از مداخله، از تمامی بیماران مرتبط موجود در یک بیمارستان گزارش گرفته شد. در همان بیمارستان، برای شش ماه بعد از مداخله، از گروه دیگری گزارش گرفته شد، و در نتیجه، گروههای همتای پیش‌آزمون و پس‌آزمون تشکیل داده شد. تحلیل و بررسی مقدار زیادی از اطلاعات مشخصات زمینه‌ای بیماران و اسناد بیمارستانی مشخص کرد تفاوتی میان دو گروه همتا وجود نداشته. این نتایج تهدید انتخاب را در مورد متغیرهای بررسی شده (اما نه برای متغیرهایی که مورد محاسبه قرار نگرفته‌اند) به حداقل می‌رساند. با این حال، بهتر بود اگر شرایط اجازه می‌داد داده‌های پیش‌آزمون و پس‌آزمون هر دو برای یک سال -و نه برای هفت ماه و شش ماه – جمع‌آوری می‌شد. چون فرایند جمع‌آوری داده صورت گرفته و داده‌های بدست آمده با اثر فصلهای مختلف (متغیر کمکی) مخلوط شده‌اند. در خصوص تهدید گذشت زمان، همکاران پژوهش اغلب روزها در بیمارستان هدف مستقر بودند، و تغییرات بی‌ربط خاصی که بتواند بهبود پس از جراحی را تحت تأثیر قرار بدهد، مشاهده نکردند. البته این هیچ تضمینی نمی‌دهد و انجام تعدیلهای لازم در طراحی بهتر از انجام محاسبه برای بی‌اثر کردن تهدیدهای روایی درونی خواهد بود. بنابراین از بیمارستانی در نزدیکی بیمارستان کنترل که در تملک همان شرکت بود و پزشکان مشترکی در آنها کار می‌کردند، نیز داده جمع‌آوری شد. داده‌های بدست‌آمده از گروه کنترل نیز این (نتیجه) را تأیید کرد که اثر مداخله به دلیل تهدید گذشت زمان نبوده است. چیزی که بواسطه این نکته آخر دستگیر ما می‌شود اهمیت دارد- یعنی همراه کردن طرح همتایان با یک عنصر طراحی مانند یک گروه کنترل دستکاری‌نشده. این دست از انوع بهبود طراحی (یعنی اضافه کردن عناصر طراحی بیشتر به منظور ارتقاء استنباط علّی) موضوعیست که در ادامه بحث به آن خواهیم پرداخت.

317 Feasible

بهبود کنترلهای طرحهای همتایان از طریق اضافه کردن پیش‌آزمون

در مطالعه‌ای که به مقایسه اثربخشی معلمین عادی و پیمانکاران خارجی(که برای افزایش موفقیت آموزشی کودکان استخدام شده بودند) می‌پرداخت، سارتسکی (Saretsky, 1972) متوجه شد که معلمین با توجه به عملکرد سالهای قبلشان، تلاش و توجه ویژه‌ای به خرج داده و عملکرد بهتری داشتند. این محقق نتایج مشاهده‌شده را به ترس معلمین از اخراج شدن (به دلیل عملکرد بهتر پیمانکاران) نسبت داد. فرض کنید برای مقاصد آموزشی، محقق میانگین نمرات بدست‌آمده از کلاسهایی که در دوره موردنظر توسط معلمین ثابت تعلیم داده شده بودند، را با نتایج سالهای گذشته همان کلاسها (که البته توسط همان معلمین مورد تعلیم قرار گرفته بودند) مقایسه کرده باشد. طرح بدست‌آمده به شیوه زیر نشان داده خواهد شد:

NR	O_1	O_2			
NR			O_3	X	O_4

O_1 و O_2 نشان‌دهنده آغاز و پایان نمرات سالیانه برای گروه همتایان قدیمی‌تری هستند که احتمال ندارد تحت تأثیر ترس معلمین برای از دست دادن شغلشان قرار گرفته باشند؛ و O_3 و O_4 نشان‌دهنده نمرات گروه بعدی است که احتمال دارد معلمین متأثر از این ترس بوده باشند. فرض صفر عبارتست از این که تغییر در یک گروه همتا برابر است با تغییرات در گروه همتای دیگر. این طرح آزمایشی را می‌توان تا هر جا که لازم است در زمان گذشته گسترش داد، به نحوی که بتوان به جای یک گروه، گروههای همتای کنترل متعددی به دست آورد. البته سارتسکی (Saretsky, 1972) داده‌های مربوط به دو سال پیش‌آزمایش را گزارش کرده است. به طور کلی، اگر مداخله در جریان است، طرح آزمایشی می‌تواند به طرف جلو (زمانهای آینده) نیز تا آنجا که لازم باشد برای سالها ادامه داد، به نحوی که بتوان تخمینهای متعددی از اثرهای مورد بررسی بدست آورد.

همانطور که در دیاگرام بالا نشان داده شده است، این طرح شبیه طرح پایه گروه کنترل غیرهم‌ارز با پیش‌آزمون و پس‌آزمون است. عمده‌ی تفاوتها این است که اولاً، اندازه‌گیری در یک دوره زمانی، زودتر در گروه کنترل اتفاق می‌افتد، و ثانیاً، فرض بر این است که همتایان در مقایسه با دیگر گروههای جفت‌شده دارای ناهمسانیهای (غیرهم‌ارزیهای) کمتری هستند. مورد دوم را می‌توان با مقایسه میانگینهای پیش‌آزمون همتایان مشاهده کرد؛ این یکی از عمده مزایای داشتن پیش‌آزمون در طرحهای همتایان است. وجود پیش‌آزمون بواسطه بوجود آوردن امکان استفاده از متغیرهای خطای۳۱۸ بین‌گروهی، توان آماری را نیز افزایش می‌دهد. پیش‌آزمون محقق را قادر

318 Error terms

می‌سازد تا به نحو بهتری تهدیدهای بلوغ و رگرسیون را ارزیابی کند، و در نتیجه، بتواند تعدیلهای آماری بهتری (البته همچنان غیرکامل) را برای ناهمسانیهای گروههای مورد مطالعه بکار بگیرد.

در این طرح آزمایشی، گذشت زمان یک تهدید جدی برای روایی درونیست. این تهدید می‌تواند شامل هر رویدادی باشد که با نتایج بدست آمده در فاصله زمانی O_3 تا O_4 همبستگی دارند، حتی اگر مجموعه‌ای از دوره‌های زمانی کنترل همتا در طرح وجود داشته باشد. تنها زمانی می‌توان امیدوار بود بتوان مشکل تهدید گذشت زمان را بی‌اثر کرد، که یک گروه کنترل غیرهم‌ارز به طرح اضافه شود، و عیناً در همان زمانهایی که در گروه آزمون همتا اندازه‌گیری صورت می‌گیرد، در گروه کنترل نیز اندازه‌گیری صورت گیرد. برخی مواقع اضافه کردن متغیرهای وابسته غیرهم‌ارز نیز می‌تواند سودمند باشد، البته اگر این کار برای موضوع تحت‌بررسی مناسب باشد.

ویرایش دیگری از این طرح، چرخه نهادی عودکننده یا متناوب نامیده می‌شود[۳۱۹]. اجرای این طرح مستلزم دسترسی داشتن به اطلاعات مدرسه و یا در اختیار داشتن حداقل دو سال زمان برای انجام مطالعه‌ای است که متضمن جمع‌آوری اطلاعات اولیه باشد. طرح مذکور به شکل زیر نشان داده می‌شود. این طرح مشتمل بر سه گروه همتاست که در، به طور مثال، سه سال تحصیلی وارد کلاس دوم می‌شوند. گروه اول مداخله و یک پس‌آزمون دریافت می‌کند؛ گروه دوم مداخله را با پیش‌آزمون و پس‌آزمون دریافت می‌کند؛ و گروه سوم مداخله دریافت نکرده و تنها مورد اندازه‌گیری قرار می‌گیرد.

NR X O_1

NR O_2 X O_3

NR O_4

در نظر داشته باشید که O_1 و O_2 ممکن است به طور همزمان انجام نشود، چون ممکن است یکی در انتهای یک سال تحصیلی باشد، و دیگری در ابتدای سال تحصیلی بعدی. این چرخه مجدداً با O_3 و O_4 تکرار می‌شود.

الگوهای خاصی از نتایج، نشان‌دهنده اثر اصلی مداخله است. این الگو عبارتست از اینکه O_1 و O_3 بالاتر از O_2 و O_4 باشد؛ O_2 متفاوت از O_4 نباشد؛ و O_1 متفاوت از O_3 نباشد. اگر علاوه بر بزرگتر بودن O_3 از O_2، O_1 نیز از O_2 بزرگتر بوده و O_3 از O_4 بزرگتر باشد، این را می‌توان تا حدودی به عنوان کنترلی برای تهدید گذشت زمان به حساب آورد. این الگو شاهدی خواهد بود از آنکه مداخله می‌توانسته در دو نقطه زمانی متفاوت اثربخش بوده باشد. اگرچه ممکن است دو نیروی تاریخی مجزا در آن واحد در حال اثرگذاری بوده باشند، و یا اینکه یک نیروی

319 Recurrent institutional cycle design

تاریخی مشابه دو بار در طول زمان رخ داده باشد۳۲۰. اما همچنان یک تهدید تاریخی واحد باید تکرار شده باشد تا بتواند هم O_1 بزرگتر از O_2 را توضیح دهد و هم O_3 بزرگتر از O_4 را. در این نوع از طرحهای همتایان، اگر افراد مشترکی در جریان مقایسه میان O_2 و O_3 شرکت داشته باشند، تهدید انتخاب نیز کاهش پیدا می‌کند.

تهدید آزمون نیز در این طرحها امکان بروز دارد، چون در بعضی مقایسه‌ها اندازه‌گیری اول با اندازه گیری دوم مقایسه می‌شود (از O_2 به O_3). از این رو، کمپبل و استنلی (Campbell & Stanley, 1963) توصیه می‌کنند گروهی که پیش‌آزمون و پس‌آزمون بر روی آن انجام می‌شود، را به طور تصادفی به دو قسمت تقسیم کنیم؛ گروهی پیش‌آزمون را دریافت کرده و گروه دیگر دریافت نکنند. تفاوت پایدار میان این دو گروه در پس‌آزمون می‌تواند به دلیل تهدید آزمون رخ داده باشد. و بالعکس، اگر میان این دو گروه تفاوت وجود نداشته باشد، می‌توان نتیجه گرفت که تهدید آزمون مسأله‌ساز نیست. در نهایت، از آنجا که تفسیر علّی به الگوی پیچیده‌ای از نتایج که در آن سه متضاد دربرگیرنده O_2 هستند، یک تغییر در افزایش O_2 می‌تواند حرفهای زیادی برای گفتن داشته باشد. در نتیجه، این طرح باید تنها با مقیاسهای اندازه‌گیری پایا و روا و همچنین با نمونه‌های بزرگ انجام شود.

بهبود طرحهای همتایان با یک متغیر وابسته غیرهم‌ارز

مینتون (Minton, 1975) برای بهبود مطالعه خود در مورد اثر فصل اول برنامه خیابان کنجد بر یادگیری کودکان کودکستانی، از یک متغیر وابسته غیرهم‌ارز استفاده کرد. او نشان داد در آن دسته از کودکان که برنامه موردنظر را تماشا کرده بودند، دانش دایره‌لغاتی که در برنامه موردنظر آموزش داده شده بود (در قیاس با لغاتی که در آن برنامه آموزش داده نشده بودند)، به طور معناداری افزایش یافته بود. این نتایج مشکل تهدید بلوغ را برطرف می‌کرد، چون دانش حروف الفبا در کودکان، عموماً در طول زمان به واسطه عوامل متعددی (مانند ارتقاء شناختی در آنها) افزایش پیدا می‌کند. اگر بلوغ تنها عاملی بود که می‌توانست نتایج را توضیح دهد، در آن صورت، انتظار می‌رفت تفاوتی میان حروفی که آموزش داده شده بود، و حروفی که آموزش داده نشده بود، وجود نداشته باشد.

طرحهایی که ترکیبی از عناصر متعدد را دربرمی‌گیرند

طی این فصل، بر این نکته تأکید شد که اضافه کردن عناصر مختلف به طراحی می‌تواند استنباط علّی را بهبود بخشد. در این بخش، سه نمونه از طرحهایی که از عناصر متعددی استفاده می‌کنند را توضیح خواهیم داد. این مثالها که به روشن کردن منطق پشت این طرحها کمک می‌کنند.

320 Reoccurred

کنترلهای جفت‌شده دستکاری‌نشده، با پیش‌آزمونها و پس‌آزمونهای متعدد، متغیرهای وابسته غیرهم‌ارز، و مداخله‌های حذف‌شده و تکرارشده

در نمونه‌ای خوب از طرحهای شبه‌آزمایشی، رینولدز و وست (Reynolds & West, 1987) اثرات کمپین «درخواست برای فروش» آریزونا، که برای فروش بلیطهای بخت‌آزمایی طراحی شده بود را موردبررسی قرار دادند. فروشگاههای شرکت‌کننده عرضه‌کننده بلیطهای بخت‌آزمایی پذیرفتند تا نوشته‌ای را برای مصرف‌کنندگان ارسال کنند که در آن نوشته شده بود: «آیا ما از شما درخواست کردیم که بلیط بخت‌آزمایی بخرید؟ اگر پاسخ منفیست شما یک بلیط مجانی دریافت می‌کنید». فروشندگان همچنین توافق کردند تا یک بلیط مجانی به مشتریانی بدهند که به خرید بلیط دعوت نشده بودند، اما خودشان درخواست بلیط کرده بودند. از آنجا که مشارکت در این کار داوطلبانه بود، گروه کنترل غیرهم‌ارز بدست‌آمده به چهار طریق (به عناصر طرح) اضافه می‌شد. اول، محققین فروشگاههای گروه آزمون را با فروشگاههای گروه کنترل از دو نظر جفت کردند، یکی اینکه گروه کنترل از همان زنجیره فروشگاهی (و در صورت امکان از همان منطقه کد پستی) انتخاب می‌شدند، و دیگر اینکه گروه کنترل و آزمون از نظر سهم بازار فروش بلیط در پیش‌آزمون جفت می‌شدند. دوم، اندازه‌گیریهای پیش‌آزمون و پس‌آزمون متعددی از طریق اندازه‌گیری میانگین فروش بلیط برای ۴ هفته قبل، و ۴ هفته بعد از شروع مداخله انجام شد. روندهای فروش پیش‌آزمون در هر دو گروه کنترل و آزمون تقریباً مشابه بود، بنابراین تفاوتها در نرخ بلوغ نمی‌توانست علّت افزایش در فروش بلیط باشد. به همین ترتیب، رگرسیون به سمت میانگین نیز غیرمحتمل به نظر می‌رسید، چون فروش گروه آزمون به طور پیوسته در طول چهار پیش‌آزمون کاهشی بود، و همچنین چون روند کاهشی فروش بلیط گروه کنترل پس از شروع مداخله نیز ادامه یافت. سوم، آیکن و وست اثرات مداخله را بر سه متغیر وابسته غیرهم‌ارزدر گروه آزمون مورد بررسی قرار دادند، و دریافتند که مداخله مورد نظر فروش بلیط را افزایش داده اما تأثیری بر فروش بنزین، سیگار، و یا خواروبار نداشته است. چهارم، این دو محقق فروشگاههایی را در نظر گرفتند که در آنها مداخله حذف، و سپس تکرار می‌شد. و یا مداخله (از نظر زمانی) بعد از دیگر فروشگاهها در آنها آغاز می‌شد، و نتایج تغییراتی که در اثر هرکدام از این اقدامات (حذف، تکرار، و شروع با تأخیر) طی این مدت در میزان فروش این فروشگاهها رخ می‌داد، را مورد مشاهده قرار دادند. در تمامی این مدت، نتایج گروه کنترل بدون تغییر بود. تقریباً تمام این تحلیلها حاکی از آن بود که مداخله «درخواست برای فروش»، فروش بلیط را بعد از آغاز برنامه افزایش داده بود، و باعث شده بود تا به سختی بتوان به گزینه‌ی دیگری (غیر از مداخله) که توضیح‌دهنده اثر مشاهده شده باشد، فکر کرد.

طرح ترکیب تکرارهای جابجاشونده با یک گروه کنترل غیرهم‌ارز

برخی اوقات محقق مداخله را بر بخشی از گروه کنترل اولیه اعمال نموده، و دیگر اعضای گروه کنترل در طول این دوره زمانی دستکاری‌نشده باقی می‌مانند. بعضی مواقع حتی محقق می‌تواند مداخله را مجدداً، و برای بار دوم

روی بخشی از گروه کنترل اجرا کند تا اثر مداخله اضافی را مورد بررسی قرار دهد. گان، ایورسون و کاتز (Gunn et al., 1985) اینکار را در مطالعه‌ای بر روی برنامه ملّی آموزش بهداشت که در ۱۰۷۱ کلاس درس اجرا شد، انجام دادند. طرح موردنظر به شکل زیر نمایش داده می‌شود. در این شکل، R نشان‌دهنده استفاده بالقوه از تخصیص تصادفی است، که مفید اما جزء ضروری لاینفک از طرح به حساب نمی‌آید.

	سال اول				سال دوم		
NR	O_1	X	O_2	R	O_3	X	O_4
				R			
NR	O_1		O_2	R	O_3	X	O_4
NR				R	O_3		O_4

ابتدا، کلاسها به دو گروه کنترل و مداخله غیرهم‌ارز تقسیم‌بندی شدند. دانش‌آموزان هر یک از گروهها از نظر میزان دانش اصول بهداشتی، قبل و بعد از سال اول اجرای برنامه، موردآزمون قرار گرفتند. سپس، گروه کنترل اولیه به طور تصادفی به دو نیم شد. برای تکرار اثر مداخله، نیمی از گروه کنترل آموزش بهداشت دریافت کردند. نیمی دیگر اما آموزشی دریافت نکردند. بعلاوه، نمونه‌ای تصادفی از گروه آزمون اولیه برای سال دوم آموزش دریافت کردند تا اثر افزایشی آموزشهای بهداشتی بیشتر موردبررسی قرار گیرد. همانطور که می‌بینید، در اینجا تکرارهای جابجاشونده به پیوستاری از کنترلهای اولیه و یک تقویت‌کننده مداخله متصل می‌شوند. این اتصال طرح تکرارهای جابجا شونده را تقویت می‌کند، علی‌الخصوص اگر در فاز دوم مطالعه، تخصیص تصادفی داشته باشیم، و یا اینکه افرادی که تقویت‌کننده مداخله را دریافت می‌کنند، در یک طرف (بالاتر یا پایین‌تر) حد جداکننده در مقیاس مربوط به نیاز برای قسمت تقویت قرار داشته باشند (یک طرح ناپیوستگی رگرسیونی. رجوع کنید به فصل ۷).

طرح یک گروه کنترل دستکاری‌نشده با دو پیش‌آزمون، و نمونه‌های وابسته و مستقل

بلک برن (Blackburn et al., 1984) و فرکوهر (Farquhar et al., 1990) همکارانشان به منظور ارتقاء مداخلات جامعه‌محور طراحی‌شده برای کاهش ریسکهای قلبی عروقی، دو پیش‌آزمون را با نمونه‌هایی که در آنها نتایج از روی نمونه‌های مستقل و وابسته اندازه‌گیری می‌شد، ترکیب کردند. در نمودار زیر منطق پشت این طرح را به تصویر می‌کشیم. با استفاده از خطهای عمودی میان Oها، نمونه‌های مستقل از یکدیگر نشان داده شده، و البته از بیان برخی پیچیدگیهای موجود در طرحهای واقعی استفاده شده در این دو مطالعه اجتناب شده است.

R	O_1	|	O_2	|		O_3	|	O_4	| O_5
R	O_1	|	O_2	|	X	O_3	|	O_4	| O_5
R	O_1		O_2			O_3		O_4	O_5
R	O_1		O_2		X	O_3		O_4	O_5

دو ردیف اول این نمودار نشان‌دهنده یک آزمایش تصادفی با جامعه‌هایی است که به شرایط کنترل و آزمون تخصیص داده شده‌اند. این آزمایش همچنین دربرمی‌گیرنده یک پیمایش پانل مقطعی است که بر روی نمونه‌هایی مستقل از خانوارهای جامعه در هر سال انجام می‌شود (اگرچه در هر مطالعه تعداد نسبتاً محدودی خانوار مورد استفاده قرار گرفتند، و بنابراین، فرضی در مورد اینکه هم‌ارزی اولیه وجود داشته است در اختیار نداریم). دو ردیف پایین در نمودار نشان‌دهنده یک پیمایش طولیست که بر روی پاسخ‌دهندگانی که برای طول مدت مطالعه دنبال شده‌اند، صورت گرفته است. نتایج اصلی مطالعه عبارت بود از اندازه‌گیری سالانه فیزیولوژیکی مشکلات قلبی، که شامل فشار خون و سطح کلسترول بود. در پیمایش مقطعی پانل، نمونه‌های تصادفی مستقل انتخاب شدند، هم به این دلیل که اندازه‌گیری‌های سالیانه فیزیولوژیکی می‌توانست پاسخ‌دهندگانی که به طور مکرر مورد اندازه‌گیری قرار می‌گرفتند را نسبت به مداخله حساس نماید، و هم به این دلیل که این مطالعه در پی تعمیم دادن نتایج به جامعه بود، و جامعه موردنظردر طول این دوره زمانی دستخوش تغییراتی شده بود. از آنجاییکه تنها سه جامعه جفت شده در مطالعه بلک برن، و دو جامعه در مطالعه فارکوهر وجود داشت، دو پیش‌آزمون برای تخمین روندهای خطی پیش از مداخله بکار گرفته شد. اگرچه، در مطالعه بلک برن، واریانس و تفاوتهای میان سالها در درون شهرها بزرگتر از میزان مورد انتظار بود، و تعدیلات آماری انجام شده در این رابطه چندان کارساز واقع نشد. بنابراین بلک برن در میانه راه طرح آزمایش را تغییر داد، بطوریکه برخی پاسخ‌دهندگان پیش‌آزمونها در چندین پس‌آزمون مورد تعقیب قرار گرفتند. در نتیجه این تغییر، نمونه‌ای طولی شکل داده شد تا مکملی باشد برای نمونه‌های مستقلی که در ادامه گرفته می‌شد. مطالعه فاکوهر از آغاز (به طور کامل) طوری طراحی شد که دربرگیرنده نمونه‌های مستقل و وابسته باشد. استفاده از عناصر طراحی متعدد و مختلف، راههای بسیاری را برای بررسی تهدیدات مختلف روایی فراهم آورد (Chaffee, Roser, & Flora, 1989). برای مثال، چافی و همکارانش تهدید گذشت زمان را از طریق مقایسه تفاوتهای میان موجهای موفقیت‌آمیز نمونه‌های مستقل در شهرهای گروه کنترل مورد بررسی قرار دادند. برای مقابله با تهدید ریزش، آنها تفاوتهای میان نمونه‌های مستقل و وابسته گروه مداخله را با تفاوتهای میان نمونه‌های گروه کنترل متناظر مقایسه کردند. اثرات ترکیبی آزمون و بلوغ از طریق مقایسه تفاوتهای میان تغییرات رخ داده در طول زمان در نمونه های مستقل (که در آن احتمال بیشتری وجود دارد تهدیدهای بلوغ و آزمون رخ دهد) با تغییرات در نمونه‌های غیرمستقل مورد بررسی قرار

گرفت (اگرچه در این موارد بلوغ کل جمعیت می‌تولند اتفاق بیافتد). هیچکدام از این راههای آزمون تهدیدهای روایی کامل و بی نقص نیستند، بلکه هر کدام تنها شواهدی پیشنهادی و نه قطعی فراهم می‌کنند.
مطالعه فارکوهر به یک دلیل دیگر نیز جالب اســـت؛ دلیلی که به هنگام بزرگ بودن واحد تخصـــیص (مانند یک جامعه و یا یک شرکت)، اهمیت می‌یابد. به دلایل هزینه‌ای و لجستیکی، به ندرت امکان داشتن چنین واحدهایی وجود دارد. مطالعه منتشـرشـده فارکوهر تنها دو واحد مداخله و دو واحد کنترل را گزارش کرد. میزان بیماریهای قلبی و عروقی در دو واحد مداخله و یک واحد کنترل (با تفاوتی اندک) کاهش یافت. اما به نظر می‌رسید علیرغم روند نزولی ملّی در طول دوره مطالعه، ریســک موردنظر در واحد کنترل دوم افزایش می‌یافت. حذف این واحد از تحلیلها تفاوت میان واحدهای مداخله و کنترل را به چیزی نزدیک به صـــفر می‌رســـاند. با در نظر گرفتن اینکه تعداد واحدهای اندازه‌گیری بســـیار اندک بودند، تلاش نداریم تا وانمود کنیم که امکان قیاس میان واحدهای مداخله و کنترل به درستی وجود داشته، اما با این وجود، واحدها قبل از تخصیص، با دقت جور شده بودند. برای مواجهه با این مشکل، نیازمند اضافه کردن واحدهای بیشتر به مطالعه (که البته کاری بسیار پرهزینه است)، و یا ترکیب مطالعات با مداخله مشابه، با یکدیگر هستیم. در مورد راه‌حل دوم (ترکیب مطالعات با مداخلههای مشابه)، مداخلهها یکســـان نیســـتند، و دیگر فاکتورهای زمینه‌ای و اندازه‌گیری به طور قطع در مطالعات مختلف متفاوت خواهد بود. هیچ دلیل متقنی برای اینکه چرا باید به تعداد واحدهای مداخله، واحد کنترل داشـــته باشـــیم وجود ندارد؛ بنابراین اضـــافه‌کردن واحدهای کنترل بیشـــتر برخی اوقات می‌تواند کم‌هزینه بوده و در عین حال بر توان آزمون بیافزاید (Kish, 1987).

عناصر طراحی

در جریان این فصل نشان دادیم که چطور حتی ضعیفترین طرحهای شبه‌آزمایشی را می‌توان با استفاده از عناصر طراحی (با دقت انتخاب‌شده) تقویت کرد. عناصری که تعداد و موجّه‌بودن تهدیدات روایی را کاهش می‌دهند. در جدول ۵.۲ این عناصـــر را به طور خلاصـــه آورده‌ایم. به طور کلی، می‌توان گفت که شـــبه‌آزمایش چیزی بیش از ترکیب اینگونه عناصـــر برای بدســـت‌آوردن تناســـب با شـــرایط خاص تحقیق نیســـت (Corrin & Cook, 1998). به

منظور راحتی بیشتر، این عناصر را در چهار گروه مورد بحث قرار خواهیم داد، که عبارتند از عناصر مرتبط با (۱) تخصیص، (۲) اندازه‌گیری، (۳)گروههای کنترل، و (۴) مداخلهها.

جدول ۵.۲: عناصر طراحی مورداستفاده در ساخت آزمایشها و شبه‌آزمایشها

تخصیص

تخصیص تصادفی

تخصیص مبتنی بر نقطه برش[۳۲۱]

دیگر تخصیصهای غیرتصادفی

جورسازی و طبقه‌بندی کردن

ماسک کردن

اندازه‌گیری

مشاهدات پس‌آزمون

پس‌آزمونهای منفرد

متغیرهای مستقل غیرهم‌ارز

پس‌آزمونهای متعدد عمده و مهم

مشاهدات پیش‌آزمون

پیش‌آزمون منفرد

پیش‌آزمون پس‌نگر

پیش‌آزمون نماینده[۳۲۲]

پیش‌آزمونهای تکرارشونده در طول زمان

پیش‌آزمونهایی روی نمونه‌های مستقل

متغیر تعدیل‌گر با تعامل پیش‌بینی‌شده

محاسبه تهدیدهای مترتب بر روایی

گروههای کنترل

گروههای منفرد غیرهم‌ارز

گروههای غیرهم‌ارز متعدد

همتایان

گروههای کنترل بیرونی در مقابل درونی

مغایرتهای برساخته[۳۲۳]

مغایرتهای برونیابی رگرسیونی

مغایرتهای نُرم‌شده

مغایرتهای داده‌های ثانویه

مداخله

تکرارهای جابجا شونده

مداخلههای معکوس

مداخلههای حذف شده

مداخلههای مکرر

تخصیص

در اغلب شبه‌آزمایش‌ها، فرایند تخصیص در اختیار پژوهشگر نیست. بلکه شرکت‌کنندگان به انتخاب خودشان به شرایط مختلف تخصیص داده می‌شوند، و یا فرد دیگری تصمیمات مرتبط با تخصیص را اتخاذ می‌کند. مانند هنگامی که پزشک تصمیم می‌گیرد که چه کسی باید تحت عمل جراحی قرار بگیرد، و یا یک معلم مدرسه و یا مدیران مدرسه تصمیم می‌گیرند که کدامیک از دانش‌آموزان و یا مدارس باید منابع جدید را دریافت کنند. شواهد قابل‌توجهی وجود دارد که نشان می‌دهد تخصیص غیرتصادفی اغلب (و نه همواره) نتایج متفاوت از آنچه تخصیص تصادفی به دست می‌دهد به بار می‌آورد (Chalmers et al, 1983; Golditz, Miller & Mosteller, 1988; Lipsey & Wilson, 1993; Mosteller, Gilbert, & McPeek, 1980; Wortman, 1992). این تفاوت زمانیکه افراد خودشان یکی از شرایط را انتخاب می‌کنند (در قیاس با زمانی که شخص سومی آنها را به شرایط مختلف تخصیص می‌دهد) پررنگتر است (Heinsman & Shadish, 1996; Shadish, Matt, Navarro, & Phillips, 2000; Shadish & Ragsdale, 1996). بنابراین تا جایی که امکان دارد باید از خودانتخابی پرهیز کرد. بعضی مواقع روشهای خاصی از تخصیص غیرتصادفی مانند تخصیص تناوبی می‌توانند به خوبی شبیه تخصیص تصادفی عمل نمایند (McAweeney & Klockars, 1998; Staines, McKendrick, Perlis, Sacks, & Deleon, 1999).

تخصیص اغلب می‌تواند به شیوه‌هایی غیر از روشهای تصادفی نیز کنترل شود. جورسازی و طبقه‌بندی هر دو می‌توانند شباهت میان گروهها را افزایش دهند. اگرچه انجام جفت‌سازی در شبه‌آزمایشها (در قیاس با انجام آن در آزمایشهای تصادفی) نیازمند دقت بسیار بیشتری است. چون زمانی که جفت‌سازی براساس مقیاسهای منفرد، غیرقابل‌اعتماد (بدون روایی و پایایی) و در یک نقطه از زمان انجام می‌شود، بیشتر از آنکه سودمند باشند، مشکل‌ساز خواهد بود. استفاده از روشهای کور که در آن مانع دید محققین، شرکت‌کنندگان، یا دیگر کارکنان

321 Cutoff
322 Proxy
323 Constructed contrasts

مرتبط با تحقیق، و خدمات تخصیص آن می‌شوند (فرایند نقاب‌زنی[۳۲۴])، می‌تواند سودمند باشد؛ البته اگر از نظر هزینه‌ای توجیه‌پذیر باشد. این فرایند می‌تواند از دو نوع سوگیری جلوگیری کند: (۱) عکس‌العمل محققین و شرکت‌کنندگان نسبت به شرایطی که به آن تخصیص داده شده‌اند (با توجه به اینکه نسبت به آن دانش دارند)؛ و (۲) تلاش افرادی که در جریان تخصیص افراد مداخله دارند، می‌تواند باعث شود نتایج یک شرایط خاص تحت تأثیر قرار گیرد. در مجموع، تمامی روشهای تخصیص غیرتصادفی مشابه یکدیگر نیستند، و تخصیص‌های غیرتصادفی می‌تولند، با اجتناب از خودانتخابی و بکارگیری کنترلهای معمول در روشهای آزمایشی- مانند جفت‌سازی و نقاب‌زنی- در زمانهایی که امکان بکارگیری آنها مقدور باشد، بهبود پیدا کند.

اندازه‌گیری

محققین می‌توانند استنباطهای علّی را از طریق کنترل ماهیت و زمانبندی اجرای مقیاسهای اندازه‌گیری مطالعه بهبود ببخشند. دلیل عمده انجام پس‌آزمون بعد از اجرای یک مداخله، از میان بردن ابهام موجود درمورد تقدم زمانی علّت بر اثر است. این تهدید زمانی بیشتر احتمال وقوع دارد که اندازه‌گیری نتایج، همزمان با اجرای مداخله انجام می‌شود. این کار در بسیاری از مطالعات همبستگی که در آنها اثربخشی مداخله و نتایج متغیرهای وابسته همگی با استفاده از یک پرسشنامه انجام می‌شود، معمول است. بدیهی است که تفکیک زمانی اندازه‌گیری این دو مؤلفه اصلی تحلیل علّی، مطلوبیت دارد. پس‌آزمون خاصی که متغیر *غیرهم‌ارز وابسته* نامیده می‌شود، نیازمند اندازه‌گیری دو سازه مرتبط (مثلاً دو مقیاس سلامتی) در پس‌آزمون است، انتظار می‌رود یکی از این دو متغیر (متغیر خروجی موردنظر) به دلیل اعمال مداخله تغییر کند. اما پیش‌بینی می‌شود متغیر دیگر (متغیر وابسته غیرهم‌ارز) به دلیل مداخله اعمال‌شده دچار تغییر نشود، اما به دیگر عوامل زمینه‌ای تهدیدکننده روایی درونی، به شیوه‌ای مشابه متغیر وابسته اصلی واکنش نشان دهد (برای مثال، هر دو متغیر به میزان مشابهی به فرایندهای بلوغی که می‌توانند بهبود دهنده شرایط سلامتی فرد باشند، پاسخ می‌دهند). اگر متغیر هدف در پاسخ به مداخله اعمال‌شده تغییر کند، اما متغیر وابسته غیرهم‌ارز تغییری ننماید، این استنباط که تغییرات مشاهده‌شده در اثر مداخله موردنظر بوده است، تقویت می‌شود. و بالعکس، اگر هر دو متغییر تغییر کنند، آنگاه این استنباط که مداخله باعث ایجاد تغییری شده است، تضعیف می‌شود، زیرا تغییرات می‌توانسته به دلیل تهدیدات روایی رخ داده باشد. استفاده از چندین پس‌آزمون عمده به محقق اجازه می‌دهد تا الگویی از شواهد مرتبط با اثرات را مورد

324 Masking

بررســـی قرار دهد. زمانی که این الگو با دانش قبلی نســـبت به الگوهای اثر نوعی یک علّت پیش‌بینی می‌شـــود، استنباطهای علّی از درجه اطمینان بالاتری برخوردار خواهند بود.

اضـــافه کردن پیش‌آزمون نیز می‌تواند به بررســـی امکان وجود ســـوگیری انتخاب و تهدید ریزش نمونه، به عنوان منبعی برای تولید اثرات مشاهده‌شده، کمک کند. انجام پیش‌آزمونهای متعدد روی یک سازه در زمانهای مناسب قبل از انجام مداخله، کمک می‌کند تا بتوانیم پرده از روندهای بلوغ، مصـنوعات رگرسـیونی، و اثرات آزمون و ابزار برداریم. بعضـــی اوقات در مواردیکه امکان جمع‌آوری داده‌های پیش‌آزمون در مورد متغیرهای نتیجه‌ای وجود ندارد، در *پیش‌آزمونهای پس‌نگر* از پاسخ‌دهندگان درخواسـت می‌شـود تا موقعیت خود در زمان پیش‌آزمون را به خاطر بیاورند. یا می‌توان *پیش‌آزمونهای نماینده* بر روی متغیرهایی که به متغیر نتیجه‌ای مرتبط هســـتند انجام داد. این عناصـر می‌توانند به روشـن شـدن احتمال وجود سـوگیری‌های ریزش و انتخاب کمک کنند، اگرچه این کمک ضـــعیفتر از کمکی خواهد بود که انجام پیش‌آزمون بر روی متغیرهای نتیجه‌ای واقعی می‌توانســـت ارائه نماید. محقق همچنین می‌تولند داده‌های پیش‌آزمون را از یک نمونه مســـتقل پیش‌آزمونی جمع‌آوری کند. این نمونه می‌تواند متشکل از شرکت‌کنندگانی باشد که متفاوت از پاسخ‌دهندگان پس‌آزمون هستند، اما فرض بر این است که مشابه آنها هستند- به طور مثال یک نمونه تصادفی از همان جمعیت.

متغیر تعدیل‌کننده اندازه و جهت اثرات مشاهده‌شده را تغییر می‌دهد. این متغیر زمانی می‌تواند به اسـتنباطهای علّی کمک نماید که محقق بتولند به طور موفقیت‌آمیزی برهم‌کنش میان یک متغیر تعدیل‌گر و مداخله را برای تولید اثرات مشـاهده‌شـده پیش‌بینی نماید. این تأیید معمولاً تهدیدهایی را برای روایی درونی دربردارد. در نهایت، محاســبه تهدیدهای روایی در ابتدای مطالعه می‌تواند به محقق کمک کند تا امکان وقوع تهدید و اینکه آیا جهت تهدید موردنظر شبیه اثر متغیر نتیجه‌ای است یا نه را تشخیص دهد.

گروههای مقایسه‌ای (کنترل)

گروههای کنترل می‌توانند اطلاعات مفیدی در مورد اسـتنباطهای نقیض (جایگزین) و اینکه در غیاب مداخله چه اتفاقی می‌افتد، ارائه نمایند. در شبه‌آزمایشها استنباطهای جایگزین اغلب وابسته به گروههای مقایسه غیرهم‌ارزی هسـتند که به دقت انتخاب شـده‌اند، به طوریکه در زمان پیش‌آزمون بیشـترین شـباهت را - از نظر مشـخصـات متعدد مشـاهده شـده یا از نظر برخی مشـخصـات خاص که محقق باور دارد می‌توانند تهدید مهمی برای روایی باشند- با گروه مداخله داشته باشند. بکار گرفتن گروههای مقایسه غیرهم‌ارز متعدد بجای داشتن تنها یک گروه مقایسـه، می‌تواند توانایی محقق را برای کشـف تهدیدهای بیشـتر و همچنین برای مثلث‌بندی‌کردن[۳۲۵] به سـمت براکت دقیق‌تری که اسـتنباط می‌شود اثر درون آن واقع شود، افزایش دهد. یکی از مقایسه‌های مفید، اسـتفاده از

325 Triangulation

گروههای کنترل همتایان است؛ گروههایی که همدوره گروه مداخله در یک سازمان (مثلاً یک مدرسه) هستند (مانند یک کلاس سوم جدید در هر سال تحصیلی). فرض بر این است که همتایان – در مقایسه با گروههای کنترل غیرهم‌ارز- قابلیت مقایسه بیشتری با یگدیگر دارند (افراد در یک سن، در یک طبقه اجتماعی- اقتصادی). مقایسه‌های غیرتصادفی با یک گروه کنترل داخلی – در مقایسه با گروههای کنترل بیرونی- می‌توانند نتایج دقیقتری بدست دهند (Aiken et al., 1998; Bell et al., 1995; Heinman & Shadish, 1996; Shadish & Ragsdale, 1996). گروههای کنترل داخلی از میان همان مجموعه شرکت‌کنندگانی انتخاب شده‌اند که گروه مداخله از آن انتخاب شده است (مثلاً از دانش‌آموزان همان مدرسه، و یا همان کلاس و یا از میان متقاضیان یک برنامه). گروههای کنترل بیرونی از مجموعه‌هایی متفاوت انتخاب می‌شوند (مانند بیمارانی در شرایط درمانی متفاوت)، و فرض بر این است که نقاط مشترک کمتری با گروه مداخله دارند. البته تعیین مرزی برای تفکیک کردن گروههای کنترل دورنی و بیرونی دشوار است، و یقیناً تمامی این گروههای مقایسه غیرهم‌ارز می‌توانند موجد سوگیریهای جدی باشند (Stewart et al., 1993).

برخی اوقات منابعی با مطلوبیت کمتر استنباطهای جایگزین را پشتیبانی می‌کنند. این منابع عبارتند از (۱) برونیابی رگرسیونی که در آن نمرات واقعی و پیش‌بینی‌شده۳۲۶ پس‌آزمون با یکدیگر مقایسه می‌شوند، (۲) یک گروه مقایسه نُرم‌شده که در آن نمرات گروه مداخله با نمونه‌های نُرم‌شده‌ای از راهنمای آزمون، و مواردی از این قبیل مقایسه می‌شود، و (۳) گروه مقایسه از داده‌های ثانویه که در آن پاسخ‌دهندگان گروه مداخله با نمونه‌هایی که از مطالعات دیگر انتخاب شده‌اند مقایسه می‌شوند. سودمندی چنین گروههای مقایسه‌ایی به این بستگی دارد که (۱) چقدر می‌توان شباهت آنها با گروه مداخله را نشان داد، (۲) آیا جفت‌سازی مفید قابل انجام است؟، و (۳) آیا امکان ساختن گروههای مقایسه متعدد وجود دارد؟ متضادهایی که در این پاراگراف از آنها یاد شد، به ندرت می‌بتوانند به نحو کافی استنباطهای جایگزین مغفول را تبیین نمایند.

مداخله

توانایی محقق برای کنترل کاربرد و برنامه زمانبندی اجرای مداخله، ابزاری قوی برای تسهیل استنباطهای علّی است. روش تکرارهای جابجاشونده اثر مداخله را در زمانی بعد، در گروهی که در ابتدا به عنوان گروه کنترل عمل می‌کرده، تکرار می‌کند. اما بهتر از آن، استفاده از گروههای کنترل متعددی است که هر کدام در زمان متفاوتی مداخله را دریافت می‌کنند. در روش مداخله معکوس مداخله‌ای اعمال می‌شود که انتظار می‌رود نتایج را (در مقایسه با نتایج موردانتظار در شرایط مداخله) معکوس نماید. در روش مداخله حذف‌شده، مداخله ابتدا اعمال می‌شود، و سپس حذف می‌شود، تا نشان داده شود که الگوی نتایج مطابق با الگوی ناشی از اعمال مداخله است.

326 projected

و در نهایت، در روش مداخله مکرر، مداخله پس از حذف‌شدن دوباره اعمال می‌شود، و این کار تا زمانی که از نظر اقتصادی قابل انجام باشد ادامه می‌یابد (برخی مواقع این طرح را طرح ABAB می‌نامند. A نشان‌دهنده مداخله، و B نشان‌دهنده حذف مداخلهست).

عناصر طراحی و شبه‌آزمایش ایده‌آل

آیا می‌توان گفت که طرح شبه‌آزمایش ایده‌آل وجود دارد؟ طرحی که عناصر مورد اشاره در این فصل را به طور بهینه ترکیب نماید؟ پاسخ معمولاً منفی است. چون بهترین طرح برای یک مطالعه، به فرضیات مورد بررسی در آن مطالعه، به ارتباط تهدیدهای مختلف با زمینه‌ای که استنباط در آن انجام می‌شود، به دانش موجود در زمینه قابل‌اعتنا بودن تهدیدهای موضوعه با توجه مطالعات قبلی، و به اینکه اضافه کردن چه عناصری از نظر اقتصادی توجیه‌پذیر است، بستگی دارد. با این وجود، اغلب شبه‌آزمایشها تعداد بسیار محدودی از عناصر بالقوه در دسترس را مورد استفاده قرار می‌دهند؛ و به زعم نویسندگان این کتاب، در مورد اغلب شبه‌آزمایشها می‌توان گفت که بهتر بود اگر توجه بیشتری به تهدیدهای مترتب بر استنباطها، و عناصر طراحی که می‌توانست موجّه بودن این تهدیدها را کاهش دهد، نشان می‌دادند.

توصیه نگارندگان در راستای نظرات فیشر (Fisher, Cited in Rosenbaum, 1984, p.41) است، که به محققین توصیه می‌کنند «نظریات خود را دقیق بیان نمایید» تا استنباطهای علّی بدست‌آمده از شبه‌آزمایشها را ارتقاء دهید (Rosenbaum, 1984, p.41). هالند (Holland, 1989) نیز به دو اصل در انجام استنباط علّی از شبه‌آزمایشها اشاره می‌کند: (۱) استنباط علّی در مطالعات غیرتصادفی در مقایسه با مطالعات تصادفی نیازمند داده‌های بیشتری هستند، و (۲) استنباطهای علّی در مطالعات غیرتصادفی در مقایسه با مطالعات تصادفی نیازمند فرضیات بیشتر در جریان تحلیل داده‌ها هستند. هالند محققین را به تأکید بیشتر بر اصل اولی (جمع‌آوری اطلاعات بیشتر) تشویق می‌نماید، زیرا جمع‌آوری اطلاعات بیشتر اغلب تنها راه آزمون فرضیات ضروری برای انجام تحلیلهای آماری بهتر است. افزودن عناصر طراحی بیشتر راهی برای جمع‌آوری داده‌های دقیقتر و متنوعتر است، که خود متعاقباً باعث بهبود استنباطهای علّی می‌شود.

نتیجه‌گیری

مطالب ارائه شده در دو فصل گذشته نشان داد که استنباطهایی علّی تولیدشده با طرحهای شبه‌آزمایشی پراستفاده، عموماً واجد حدی از ابهام هستند. بر این اساس، کاربران این طرحها باید برای تحمل این ابهام آماده بوده و فرض کنند که دیگر توضیحات علّی جایگزین قابل‌اغماض هستند، و یا اینکه از طرح قویتری استفاده کنند. در فصل آینده همین رویه را ادامه خواهیم داد. سریهای زمانی منقطع[۳۲۷] به طور اخص ساختاری

327 Interrupted time series

مستحکم برای پشتیبانی از استنباطهای علّی تأمین می‌کنند. اما هنگامی که مؤلفه‌های طراحی مورد اشاره در این فصل به سریهای زمانی منقطع اضافه می‌شوند، شبه‌آزمایشی حاصل می‌شود که نتایج استنباطی آن برخی اوقات قابل رقابت با نتایج بدست آمده از آزمایشات تصادفی است.

پیوست ۵.۱: پیشرفتهای مهم در تحلیل داده‌های حاصل از طرحهای دارای گروههای غیرهم‌ارز

آماردانها و اقتصاددانان اخیراً توجه ویژه‌ای به تحلیل داده‌های بدست آمده از طرح‌های دارای گروههای غیرهم‌ارز معطوف داشته‌اند. بخش اعظمی از یافته‌های آنها تا حد زیادی آماری و خارج از بحث این کتاب است. اما با توجه به اینکه فصل حاضر به شبه‌آزمایشها می‌پردازد، اگر این پیشرفتهای متأخر معرفی نشود، چیزی کم خواهد بود. پیشرفتهایی که امید می‌رود نه تنها به عنوان جایگزینی برای بخش اعظمی از شبه‌آزمایشها عمل کنند، بلکه بتوانند به عنوان نیرویی مکمل در طرحها برای مقابله با سوگیریهایی که بهترین طرحها نیز تا به حال از حل آنها عاجز مانده‌اند، کمک نمایند. از این منظر، شعار اصلی عبارتست از اینکه «تعدیلات آماری تنها زمانی انجام می‌شوند که بهترین کنترلهای طراحی ممکن بکار گرفته شده باشند». وینشیپ و مورگان (Winship & Morgan, 1999) در مقاله خود مروری کامل از مطالب مرتبط با این موضوع ارائه می‌کنند.

نمرات تمایل[۳۲۸] و سوگیری های مخفی

در طول قرن بیستم، آماردانها آزمایشهای تصادفی را ترجیح داده، و در نتیجه کمتر به شبه‌آزمایشها توجه کرده‌اند (Shadish & Cook, 1999). این ترجیح تا حدی به دلیل تعامل‌پذیری[۳۲۹] سوگیری انتخاب است، زیرا زمانی که فرایندهای زمینه‌ای و ماهیت آنها نامعلوم است، ساختن مدلهای آماری بسیار دشوار است. گرچه برخی از آماردانها این مشکلات را مورد مطالعه قرار داده و نتایج مفیدی نیز بدست آورده‌اند (Holland, 1986; Rosenbaum, 1984, 1995a; Rubin, 1974, 1994). روزنباوم (Rosenbaum, 1995a) بسیاری از این نتایج را در مقاله خود به طور خلاصه مطرح کرده، و به عنوان نمونه مثالهای کاربردی فراوانی در حوزههای اپیدمیولوژی (C. Drake & Fisher, 1995)، پزشکی (Connors et al., 1996; Smith, 1997; Stone et al., 1995)، ارزیابی برنامه‌های آموزش ضمن خدمت (Dehejia & Wahba, 1999)، و آموزش دبیرستانی (Rosenbaum, 1986) ارائه داده است.

یکی از پیشرفتهای سودمند حاصل شده، نمرات تمایل هستند. این نمره عبارت است از احتمال پیش‌بینی‌شده در یک معادله رگرسیون لجیستسک در مورد بودن در گروه مداخله (در مقابل بودن در گروه کنترل). محاسبه دقیق پیش‌بینهای احتمالی انتخاب برای گروهها می‌تواند صحت نمرات تمایل را افزایش دهد. هدف، لحاظ کردن تمامی متغیرهایی است که در فرایند انتخاب نقش دارند. این شامل برهم‌کنشها و دیگر مؤلفه‌های

328 Propensity
329 Intractability

غیرخطی(Rosenbaum& Rubin, 1984; Rubin & Thomas, 1996)، و متغیرهایی که احتمالاً با نتایج مرتبط هستند (حتی اگر این ارتباط ضعیف بوده باشد) می‌شود (Rubin, 1997). فقط درحالتی که بتوان متغیری را به دلیل وجود این اجماع که متغیر مربوطه ارتباطی با نتایج ندارد، و یا متغیر تصادفی کمکی[330] مناسبی نیست، مستثنی کرد. در غیر این صورت، توصیه می‌شود متغیر موردنظر در مدل نمرات تمایل لحاظ شود، حتی اگر از نظر آماری معنادار نباشد (Rubin & Thomas, 1996, p.253). برخی نویسندگان همچنین توصیه می‌کنند، در صورتی که لندازه نمونه اجازه می‌دهد، هر متغیر پیش‌آزمونی که میان گروههای غیرهم‌ارز (در نرخ خطای نوع اول بالاتر از معمول) متفاوت است، به عنوان متغیر پیش‌بین در نظر گرفته شود (Canner & Wahba, 1999; C. Drake, 1993). متغیرهای پیش‌بین نباید بواسطه مداخله ایجاد شده باشند. فهمیدن این موضوع نیازمند انجام اندازه‌گیری قبل از آغاز مداخله است. داده‌ها نشان می‌دهند که درست مدل‌کردن فرم رگرسیون (لحاظ کردن درست متغیرهای برهم‌کنشی و غیرخطی در مدل)، کم‌اهمیت‌تر از لحاظ کردن تمامی پیش‌بینهای مرتبط با عضویت در گروه است (Dehejia & Wahba, 1999; C. Drake, 1993). در مورد مداخلههای متعدد، نمرات تمایل را می‌توان به طور مجزا برای هر مقایسه دوتایی محاسبه کرد (Rubin, 1997).

رگرسیون لجیستیک مجموعه متغیرهای کمکی تصادفی مربوط به هر شرکت‌کننده را به یک نمره تمایل تقلیل می‌دهد. در نتیجه، امکان جفت‌کردن و طبقه‌بندی کردن بر اساس متغیرهای متعدد ضروری را به طور همزمان فراهم می‌آورد. می‌توان از جفت‌سازی استاندارد استفاده نمود که در آن، یک گروه مداخله با یک گروه کنترل جور می‌شوند. اما روزنباوم (a۱۹۹۵) نشان می‌دهد که اینگونه جفت‌کردنها معمولاً فاصله میان گروههای درون هر طبقه را از نظر نمرات تمایل به حداقل نمی‌رساند. او در عوض جفت‌سازی بهینه را پیشنهاد می‌کند، در این نوع جفت‌سازی، هر زیرمجموعه مشتمل بر (۱) یک شرکت‌کننده دستکاری‌شده منفرد، و یک یا بیشتر شرکت‌کننده کنترل است؛ یا (۲) یک شرکت‌کننده کنترل منفرد، و یک یا بیشتر شرکت‌کنندگان دستکاری‌شده. جفت‌سازی بهینه الگوریتمی را برای به حداقل رساندن مجموع تفاوتهای نمونه‌ها در شرایط کنترل و مداخله از نظر نمرات تمایل بکار می‌گیرد. این روش امکان حذف جفتهای قبلی برای ایجاد جفتهای جدید را فراهم می‌آورد، اگر آن رویه بتواند حداقل مجموع تفاوتها را درطول شرایط بدست آورد (Rosenbaum, 1995a). برگسترال و همکارانش (Bergstralh et al., 1996) یک ماکرو SAS برای جفت‌کردن بهینه و ایسرمن و ریفان (Isserman & Rephann, 1995) نمونه‌ای از کاربردهای آن را در علوم اجتماعی ارائه می‌کنند (Dehejia & Wahba, 1999; Gu & Rosenbaum, 1993; Heckman, Ichimura, & Todd, 1997’ Marsh, 1998). برای مثال، یک نفر می‌تواند از نظر یک متغیر جفت شود، در حالی که به طور همزمان، در دیگر متغیرها –مانند سن و جنسیت- نیز

330 Covariate

جفت می‌شـود. اگرچه تا به حال مطالعه جامعی در مورد مزایا و نقاط ضـعف انواع گزینه‌های جفت‌سـازی انجام نشده است.

کوکران (Cochran, 1968) نشان می‌دهد که اگر طبقه‌بندی انجام شود، پنج زیرمجموعه اغلب برای حذف بیش از ۹۰٪ از سوگیریهای ناشی از تقسیم‌بندی متغیر یا متغیرکمکی تصادفی کفایت می‌کند (Rosenbaum & Rubin, 1984, p.516). بنابراین عموماً پنج طبقه ساخته می‌شود که دربرگیرنده تمامی موارد مداخله و کنترلیست که در محدوده نمرات تمایل مشـابه قرار دارند. این طبقه‌بندی تحت تأثیر نقض شـرط خطی بودن قرار نگرفته، و گروههای کنترل و مداخله را متوازن می‌سازد، به گونه‌ای که درون هر طبقه از نظر نمرات تمایل یکدست هستند. بنابراین، اگر نمرات تمایل طبقه‌ای خوب عمل کرده باشـد، تفاوتهای میان شـرکت‌کنندگان مداخله و کنترل در متغیرهای پیش‌بین به دلیل شـانس خواهد بود. سـپس میانگین گروه مداخله به صـورت تخمینی از یک میانگین وزنی از میانگینهای طبقه‌ای پنج گروه مداخله محاسـبه می‌شـود؛ میانگین گروه کنترل نیز به همین صـورت تخمین زده می‌شـود. رابینز و همکارانش (Robins et al., 1999) روشـی را به عنوان جایگزین برای وزن‌دهی پیشـنهاد می‌کنند که به تمایل دریافت مداخله در واقع دریافت‌شـده وابسـته اسـت. مداخله‌ای که می‌توانسـته مزایایی داشـته باشـد، خصـوصـاً برای مداخلههای متفاوت از نظر زمانی. محقق باید بررسـی کند که به چه میزان طبقه‌بندی بر اساس نمرات تمایل در تعدیل تفاوتها در متغیرهای کمکی تصادفی مشاهده‌شده خوب عمل کرده اسـت. این کار را از طریق وارد کردن هر متغیر کمکی تصـادفی (به طور مجزا) و نمره تمایل در یک تحلیل واریانس ۲ (مداخله)X ۵ (طبقه) انجام می‌دهند. وجود برهم‌کنش معنادار نشان‌دهنده آن است که نمرات تمایل برای متغیرهای کمکی تصادفی به خوبی تعدیل نشده‌اند. این وضعیت زمانی بیشتر احتمال وقوع دارد که دو گروه از نظر متغیرهای کمکی پیش‌آزمون، به میزان بیشـتری تفاوت داشـته باشـند. برخی مواقع این مشـکل می‌تواند با اضافه‌کردن متغیرهای غیرخطی به معادله نمرات تمایل بهبود داده شود.

در نهایت، نمرات تمایل را می‌توان به عنوان متغیر کمکی در ANCOVA به کار گرفت. هنگامی که فرضیات معمول ANCOVA برقرار بوده، و مدل دقیقاً صحیح است (مثلاً توانسته منحنی‌الخطی را به درستی در مدل لحاظ نماید)، تعدیل متغیرکمکی کارآمدتر از جفت‌کردن و طبقه‌بندی‌کردن اسـت. اگر مدل چندان دقیق نباشـد، تعدیل متغیرکمکی کمتر می‌تواند سـوگیریها را کاهش دهد، و یا ممکن اسـت حتی آنها را افزایش دهد. برخی محققین در این تردید دارند که مدلهای کوواریانس تا چه اندازه می‌توانند شکل کارکردی درست را مدل نمایند (مثلاً H. White, 1981; Winship & Morgan, 1999). دهجیا و واهبا (Dehejia & Wahba, 1999) دریافتند که اگر آزمایش تصادفی را معیار[۳۳۱] قرار دهیم، جفت‌سازی بهتر از کوواریانس عمل می‌کند، حتی اگر برخی متغیرهای غیرخطی به مدل کوواریانس اضافه شوند. خوشبختانه جفت‌سازی و طبقه‌بندی بر اساس نمرات تمایل را می‌توان به طور

331 Benchmark

ترکیبی در تحلیلهای کوواریانس بعدی بکار گرفت. نتایج کاراتر و مستحکم‌تر از زمانی است که هر کدام از آنها به طور منفرد به کار گرفته می‌شوند. این ANCOVA می‌تواند شامل [آن دسته] متغیرهای پیش‌بینی مرتبط با عضویت در گروه باشد، که برای محاسبه نمرات تمایل بکار گرفته می‌شوند (Rubin & Thomas, 2000; Stone et al., 1995). یک متغیر پیش‌بین می‌تواند هم نشان‌دهنده واریانس در عضویت گروه باشد، و هم واریانس در نتایج. اگرچه این حالت دوم چندان معمول نیست، یک پیش‌بین می‌تواند مسئول واریانس در عضویت گروهی، و واریانس در خروجیها باشد. تا جایی که منابع واریانس، متعامد[۳۳۲] باشد (غالباً یک سوال عملی در رابطه با هر مورد)، لحاظ کردن متغیر پیش‌بین در معادله نتایج نهایی می‌تواند کارایی را افزایش داده، و سوگیری تخمینهای نهایی را کاهش دهد.

چهار محدودیت، هیجان و علاقه موجود نسبت به نمرات تمایل را خنثی می‌کند. اول، این نمرات زمانی به بهترین نحو عمل می‌کنند که اندازه نمونه بزرگ باشد، اما در بسیاری از شبه‌آزمایشها نمونه‌ها کوچک هستند. دوم، محققین باید همپوشانی میان شرایط مختلف را از نظر نمرات تمایل مورد بررسی قرار دهند. زمانی که همپوشانی بسیار ناچیز است، امکان داشتن طبقه‌ها و یا جفت‌های متعدد از افرادی از شرایط مداخله‌های متضاد وجود ندارد، که این باعث می‌شود تا اندازه نمونه، تعمیم‌پذیری و صحت نتایج علّی تا حد زیادی محدود شود. سوم، روشهای محاسبه نمرات تمایل برای مواقعی که متغیرهای پیش‌بین دارای داده گمشده[۳۳۳] هستند معرفی شده‌اند، و این موضوع در عمل بسیار حیاتی است چون وجود داده‌های گمشده بسیار معمول است. چهارم، این روش فرض را بر این می‌گذارد که هیچ متغیر کمکی تصادفی دیگری که با نتایج همبستگی داشته و بتواند تمایل به بودن در یک گروه را تحت تأثیر قرار دهد، وجود ندارد. این یک پیشفرض قوی‌ست. تخصیص تصادفی همانطور که انتظار می‌رود، مداخلهها را از نظر متغیرهای هم‌تغییر[۳۳۴] مشاهده‌شده و مشاهده‌نشده متوازن می‌سازد. اما تعدیلات نمرات تمایل، تنها مداخله‌ها را از نظر کوواریانسهای مشاهده‌شده متوازن کرده، و سوگیریهای پنهان بوجود آمده ناشی از متغیرهای مشاهده‌نشده را دربرنمی‌گیرد. این تعدیلات در صورتی می‌توانند سوگیری‌های پنهان را کاهش دهند که نمرات تمایل از متغیرهای پیش‌بین عضویت در گروه و متغیرهای خروجی متعددی (تا آنجا که توجیه‌پذیر است) ساخته شده باشند. اگرچه، به ندرت امکان دارد که تمامی چنین متغیرهایی را بشناسیم. از آن گذشته، هزینه و محدودیتهای لجستیکی اغلب مانع از آن می‌شود که محققین متغیرهایی که

332 Orthogonal
333 Missing
334 Covariate

امکان اثرگذاری داشته باشند را در محاسبات لحاظ نمایند. بنابراین، حتی زمانی که بهترین تحلیلهای تمایل نیز انجام می‌شود، همچنان سوگیریهای پنهان در تخمینهای حاصل از شبه‌آزمایشها باقی می‌ماند.

تحلیل حساسیت، به عنوان دومین نوع از پیشرفتهای صورت‌گرفته در آمارها به طور مستقیم از احتمال وجود این سوگیریهای پنهان نشأت می‌گیرد. تحلیلهای حساسیت برای ارزیابی اینکه آیا سوگیریهای پنهان در هر اندازه، نتایج مطالعه را تغییر خواهند داد یا نه، به کار گرفته می‌شوند. این تحلیلها تعیین خواهند کرد که چه میزان سوگیری پنهان لازم است وجود داشته باشد تا نتایج تغییر کند. عموماً این تحلیل را در قالب طیفی از تفاوت معنادار میان گروهها تا عدم وجود تفاوت و یا بالعکس ارایه می‌کند (Gastwirth, 1992; Gastwirth, Krieger, & Rosenbaum, 1994; S. Greenhouse, 1982; Marcus, 1997b; Psaty et al., 1999; Rosenbaum, 1986).

اخیراً، تحقیقات مشابهی در زمینه اقتصادسنجی صورت گرفته که در قسمت بعد به آنها خواهیم پرداخت (Manski, 1990; Manski & Nagin, 1998).

در ادامه، تحقیقات صورت گرفته توسط روزنباوم در حوزه تحلیل حساسیت، مورد بررسی قرار می‌گیرد. در یک آزمایش تصادفی که تخصیص تصادفی ساده را بکار گرفته است، احتمال تخصیص یافتن به گروه آزمایشی و کنترل برابر است؛ در نتیجه احتمال تخصیص یافتن به گروه مداخله برابر است با ۰/۵. در این حالت، سطح معناداری (نرخ خطای نوع اول) بدست‌آمده در آزمون آماری تفاوت میان دو گروه درست و دقیق است. در آزمایشهای غیرتصادفی اما، این احتمالها می‌تواند از ۰/۵ متفاوت باشد. برای مثال، ممکن است مردان بیشتر از زنان برای مداخله‌های مرتبط با آموزشهای ضمن خدمت انتخاب شوند. از آنجا که این احتمالها متفاوت از ۵۰/۵۰ است، سطح معناداری بدست‌آمده در آزمون آماری تفاوت میان گروهها، می‌تواند از صحت کمتری برخوردار باشد (اگر فرض نماییم که متغیرهای حذف‌شده منجر به سوگیری، با نتایج در ارتباط هستند). متأسفانه بدون شناخت این سوگیریهای پنهان که می‌توانند موجب تغییرات در تخصیصها شوند، نمی‌توانیم تشخیص دهیم که آیا سطوح معناداری خیلی پایین است یا خیلی بالا. تحلیل حساسیت تعیین می‌کند که بالاترین و پایینترین سطوح معناداری ممکن، به چه اندازه از آنچه که در آزمایشهای تصادفی بدست آمده است متفاوت خواهد بود. در تحلیل حساسیت، اینکار برای فرضیات متفاوت در مورد اینکه احتمال تخصیص‌داده‌شدن به گروهها تا چه اندازه از ۰/۵ متفاوت است، به طور مجزا انجام می‌گیرد. این تحلیل می‌تواند اطلاعات آسیب‌شناختی مهمی را درباره میزان سوگیری تخصیص در متغیری مرتبط با نتایج، که می‌تواند معناداری نتایج را تغییر دهد، ارائه کند.

روزنباوم (۱۹۹۱a) مثالی را ارائه می‌کند که در آن، سطح معناداری مشاهده‌شده در شبه‌آزمایش عبارت است از P= ۰/۰۰۵۷ و این نشان‌دهنده اثربخش‌بودن مداخله به حساب می‌آمد. تحلیل حساسیت بر روی داده‌های خام نشان داد که دامنه[۳۳۵] معناداری‌های ممکن از حداقل ۰/۰۰۰۴ تا حداکثر ۰/۰۳۶۷ باشد. زمانی که احتمال

335 Range

تخصیص دامنه‌ای از ۰/۴ تا ۰/۶ داشته باشد، هر دو این حداکثر و حداقل، این نتیجه‌گیری را که مداخله موثر بوده است را تقویت می‌کند. اگرچه، دامنه باریک و محدود احتمال تخصیص (۴۰/۶۰)، نشان‌دهنده افتراق ناچیز از تخصیص تصادفی به دلیل سوگیریهای پنهان است. اگر احتمال تخصیص یافتن به شرایط مختلف دامنه‌ای از ۰/۲۵ تا ۰/۷۵ داشته باشد، و حداقل سطح معناداری کوچکتر از ۰/۰۰۰۱ باشد، اما حداکثر ۰/۲۴۲۰ باشد، این حداقل و حداکثر نشان می‌دهد که اثر معناداری وجود ندارد. احتمالها نشان می‌دهد که اگر متغیرهای محاسبه‌نشده‌ای وجود داشته باشند که می‌تواند تخصیص اعضاء به شرایط مختلف را تحت تأثیر قرار دهند، به‌گونه‌ای که برخی افراد شانس بیشتری به میزان نسبت ۳ به ۱ (یعنی ۰/۷۵ به ۰/۲۵) از دیگر افراد برای تخصیص‌داده‌شدن به گروه مداخله داشته باشند، در اینصورت، سوگیری پنهان می‌توانسته اثر مداخله نادرستی ایجاد کرده باشد، در جایی که در حقیقت اثری وجود نداشته است (یا اینکه این اثر باعث دیده‌نشده اثرات مداخله بزرگتری شده باشد). الگوی سطوح معناداری ماکسیمم و مینیمم و عدم‌تشابه‌ها در احتمالهای تخصیص لازم برای تولید این سطوح، در مطالعات مختلف متفاوت است. برخی مطالعات تنها نسبت به فرضیات مرتبط با سوگیریهای پنهان بسیار حاد، آسیب‌پذیر هستند؛ اما برخی دیگر به نظر می‌رسد نسبت به تمامی انواع سوگیری آسیب‌پذیر باشند. تحلیلهای حساسیت در واقع مشخص نمی‌کنند که آیا سوگیری وجود دارد یا نه. بلکه تنها می‌توانند نشان‌دهند که آیا مطالعه به درجات متفاوت نسبت به سوگیریها آسیب‌پذیر است یا خیر. روزنباوم (a۱۹۹۱) مطالعه‌ای را مورد بحث قرار می‌دهد که نسبت به سوگیریهای پنهانی که موجب شده‌اند احتمال تخصیص، دامنه‌ای از ۰/۰۹ تا ۰/۹۱ داشته باشد، آسیب‌ناپذیر است. اما بررسی‌های بعدی نشان داد که احتمالاً سوگیری بزرگتری در مطالعه وجود داشته است. تشخیص واقعی سوگیری پنهان در یک مطالعه به راحتی قابل دستیابی نیست. اما برخی مواقع عناصر طراحی که در این فصل و فصل قبل مورد بحث قرار دادیم – مانند استفاده از متغیرهای وابسته غیرهم‌ارز و یا گروههای کنترلی که دارای عملکردهای مشخصی در زمینه برخی متغیرهای هم‌تغییر مشاهده‌نشده هستند – می‌توانند سودمند باشند. برای مثال، دهجیا و واهبا (۱۹۹۹) پیشنهاد می‌کند زمانی که تعدیلات (اصلاحات) نمرات تمایل در مورد گروههای مقایسه غیرهم‌ارز متعدد نتایج بسیار متغیری بدست آورد، احتمال وجود سوگیری پنهان وجود دارد.

زمانی که تحلیلهای حساسیت با جفت‌سازی در نمرات تمایل ترکیب می‌شود، این ابزار تحلیلی مهم را در مجموعه ابزارهای شبه‌آزمایش مهیا می‌نماید. ما امیدوار هستیم این ابزارها به طور گسترده‌تری مورد استفاده قرار بگیرند تا بتوان تجربیات کاربردی بیشتری درباره صحت و به صرفه بودن آنها بدست آورد.

مدلسازی سوگیری انتخاب[۳۳۶]

336 Selection bias

با توجه به اینکه تحلیل نمره تمایل را نمی‌توان برای سوگیریهای پنهان تعدیل کرد، و اینکه تحلیل‌های حساسیت نمی‌توانند نشان‌دهند که آیا چنین سوگیری‌هایی وجود دارند یا نه، وجود روشی که این ضعف‌ها را اصلاح کند، مطلوب خواهد بود. در ۲۵ سال گذشته، اقتصاددان متعددی، از جمله هکمان (مثلاً Barnow, Cain & Goldberger, 1980؛ Cronbach, Rogosa, Floden & Price, 1977؛ Director, 1979؛ W.Greene, 1985, 1999؛ Heckman, 1979؛ Heckman & Hotz, 1989a, 1989b؛ Heckman, Hotz & Dabos, 1987؛ Heckman & Robb, ۳۳۷ Stromsdorfer & Farkas, 1980؛ 1985, 1986a)، رویکردهای مختلفی را معرفی کرده‌اند، با این امید که سوگیری انتخاب بین گروه‌های غیرهم‌ارز را تعدیل کنند، تا بتوانند برآوردی بدون سوگیری، از آثار مداخله بدست آورند. این روش‌ها، از نظر آماری پیچیده هستند و افراد فاقد دانش آماری پیشرفته معمولاً نمی‌توانند آنها را به سهولت اجرا کنند. این روش‌ها، شامل خانواده‌ای از مدل‌ها هستند که در آنها، پیشفرضهای مختلفی در خصوص انتخاب صورت می‌گیرد. به منظور مطالعه بیشتر این روشها می‌توان به منابعی همچون آخن (۱۹۸۶)، فاستر و مک لاناهان (۱۹۹۶)، موتیف (۱۹۹۱) و خصوصاً وینشیپ و مورگان (۱۹۹۹) اشاره داشت (مانند Achen, 1986؛ Foster & Mclanahan, 1996؛ Moffitt, 1991؛ Newhouse & McClellan, 1998؛ Rindskorf, 1986؛ Wiship & Mare, 1992 و به ویژه Winship & Morga, 1999).

در مدل سوگیری انتخاب ساده، از دو معادله استفاده می‌شود، که عبارتند از معادله انتخاب، و معادله خروجی یا نتایج. همانند مدل‌های نمره تمایل، در اینجا نیز، معادله انتخاب، عضویت واقعی در گروه را از روی مجموعه قیدهای پیش‌فرض انتخاب، تحت شرایط معین پیش‌بینی می‌کند، و یک نمره عضویت گروه پیش‌بینی‌شده بدست می‌دهد. این نمره را می‌توان در معادله خروجی، به جای متغیر مجازی[۳۳۸] مداخله قرار داد، یا در کنار متغیر مجازی، در معادله وارد کرد. اگر معادله انتخاب بتواند عضویت گروه را تقریباً به صورت کامل پیش‌بینی کند، و اگر پیشفرضهای دیگر مانند نرمال بودن مشاهدات برآورده شوند، آنگاه، ضریب وابسته به متغیر مداخله مجازی پیش‌بینی‌شده در معادله برآورد اثر، یک برآورد بدون سوگیری از اثر مداخله خواهد بود. بر خلاف روش‌های نمره تمایل، در مدل‌های سوگیری انتخاب، وجود همبستگی میان خطاها در معادله انتخاب و معادله

۳۳۷ Wainer (۱۹۸۶، صفحات ۵۷ – ۶۲، ۱۰۸ تا ۱۱۳) بحث مفهومی جان تاکی، حان هرینگان و جیمز هکمان در نسخه‌های ۱۹۸۵ و ۱۹۸۶ از مقالات همکان و راب را دوباره منتشر کرده است.

338 Dummy variable

خروجی امکان‌پذیر است. این همبستگی با این فرض بدست می‌آید که رابطه دومتغیره بین خطاها، به صورت نرمال دومتغیره در نظر گرفته می‌شود.

این مدل‌ها ارتباط نزدیکی با طرح‌های ناپیوستگی رگرسیون دارند. یعنی با داشتن دانش کامل نسبت به مدل انتخابی که بر مبنای آن، شرکت‌کنندگان به شرایط تخصیص داده شده‌اند، و وارد کردن مستقیم متغیر برش[۳۳۹] در مدل برآورد اثرات، برآوردی بدون سوگیری از آثار مداخله انجام می‌دهند. ناپیوستگی رگرسیونی، نیازی به یک معادله انتخاب ندارد، زیرا طراحی خود منجر به پیش‌بینی بی‌نقص انتخاب براساس نمره برش می‌شود. در مدل‌های سوگیری انتخاب نیز به شکلی مشابه، اگر باقیمانده معادله انتخاب، تفاوت زیادی با صفر داشته باشد، (که مانند این است که بگوییم عضویت گروه پیش‌بینی شده، انطباق خوبی با عضویت واقعی در گروه ندارد)، آنگاه مدل سوگیری انتخاب نمی‌تواند برآوردهای بدون‌سوگیری از اثرات مداخله بدست دهد. این مسأله زمانی رخ می‌دهد که متغیری که باعث بهبود پیش‌بینی عضویت گروه شده و با خروجی مرتبط است، از معادله انتخاب حذف شود. این حذف، سبب همبستگی بین جملات خطا و پیش‌بین‌ها در معادلات انتخاب و برآورد اثر می‌شود، که خود دلیل برآورد سوگیرانه اثرات مداخله خواهد بود. شکل تابعی معادله انتخاب نیز همانند ناپیوستگی رگرسیون باید به درستی مشخص گردد. به عنوان مثال، اگر جملات غیرخطی یا برهم‌کنشی که بر عضویت گروه تأثیرگذار هستند، حذف شوند، معادله برآورد اثرات ممکن است برآوردهای سوگیرانه‌ای بدست دهد.

مدل‌های سوگیری انتخاب به صورت گسترده مطالعه شده، و مزایا و معایب آنها مورد بررسی قرار گرفته است[۳۴۰]. از جمله جنبه‌های مثبت مدلهای انتخاب این است که به جای اینکه تنها مطابق با متغیرهای کمکی مشاهده‌شده تعدیلات انجام دهند، سوگیری پنهان را در نظر می‌گیرند و می توانند برخی داده‌های عملی و کاربردی را به شکل مثبتی تفسیر کنند (Heckman & Hotz, 1989a؛ Heckman, Hotz, & Dabos, 1987؛ Heckman & Todd, 1996؛ Reynolds & Temple, 1995)، و در نتیجه اشتیاق محققان را برای کار بیشتر و توسعه این مدل‌ها بربیانگیزند (Moffitt, 1989). منتقدین اما به حساسیت بالای این مدلها نسبت به نقض پیشفرضها اشاره می‌کنند، و بسیاری از آماردانان نیز به این مدل‌ها با دیده تردید می‌نگرند (مانند Holland, 1989؛ Little, 1985؛ Wainer, 1986). علاوه بر این، برخی مطالعات نشان می‌دهند که این مدل‌ها، ناتوان از تولید نتایجی مشابه[۳۴۱] نتایج آزمایش‌های تصادفی هستند. به عنوان مثال، لالونده و مینارد (Lalonde & Maynard, 1987) نتایج بدست آمده از آزمایش تصادفی را با نتایج تحلیل سوگیری انتخاب همان داده‌ها که به کمک یک کنترل شبه‌آزمایشی بدست آمد، مقایسه کردند، و دریافتند که این دو جواب انطباق خوبی با یکدیگر ندارند. پیش‌فرض این است که آزمایش

339 Cutoff

۳۴۰ برای دیدن یک دیدگاه جدید در این باره از دیدگاه جامعه شناسی علم، به مطالعه بریسلاو (Breslau, 1997) مراجعه کنید.

341 Well-approximate

تصادفی صحیح است. مطالعات مربوطه اما، نتایج نویدبخشی را بدست نیاوردند (Fraker & Maynard, 1986, 1987؛ Friedlander & Robins, 1995؛ Lalonde, 1986؛ Murnane, Newstead & Olsen, 1985, Newstead & Olsen, Virdin, 1993؛ Stolzenberg & Relles, 1990)[۳۴۲]. بنابراین، حتی برخی اقتصاددانان آزمایش‌های تصادفی را به آزمایش‌های غیرتصادفی که در آنها از مدل‌های سوگیری انتخاب بهره گرفته می‌شود، ترجیح می‌دهند (Ashenfelter & Card, 1985؛ Barnow 1987؛ Burtless, 1995؛ Hollister & Hill, 1995). مدافعان پاسخ می‌دهند که در برخی از این مطالعات، از داده‌هایی استفاده شده که معیارهای لازم برای آنکه مدل بتواند بدرستی عمل نماید را برآورده نمی‌سازند. به عنوان مثال، همکان و هوتز (Heckman & Hotz, 1989a, 1989b؛ Heckman, Hotz & Dahos, 1987) پیشنهاد می‌کنند که یک مدل معتبر سوگیری انتخاب نباید در پیش‌آزمون اختلافی میان شرکت‌کنندگان و کنترل‌ها پیدا کند. همچنین، در پس‌آزمون نباید بین کنترل‌های تصادفی و غیرتصادفی تفاوتی وجود داشته باشد (البته، اگر کسی گروهای کنترل تصادفی داشته باشد، استفاده از برآورد سوگیری انتخاب چندان جذابیتی نخواهد داشت). اما حتی زمانی که این شروط برآورده می‌شوند نیز، باز هم دقت برآوردهای حاصله نگران‌کننده است (Friedlander & Robins, 1995).

تلاش برای ساختن مدل‌های سوگیری انتخاب بهتر، همچنان ادامه دارد (Heckman & Roselius, 1994, 1995؛ Heckman & Todd, 1996). بل و همکارانش (Bell et al., 1995) بر این باورند که رویدادهای دهه ۱۹۷۰ منجر به بکارگیری بیشتر گروه‌های کنترل خارجی (از جمله آنهایی که از آرشیوهای پیمایشهای ملی در مدل‌های سوگیری انتخاب استخراج شده‌اند)، و استفاده کمتر از گروه‌های کنترل داخلی شد. امروزه، مجدداً توجه به گروه‌های کنترل داخلی گسترش پیدا کرده است، با این فرض که این گروه‌ها می‌توانند نسبت به کنترل‌های خارجی، شباهت بسیار بیشتری به گروه مداخله داشته باشند. اگرچه، این مسأله برای مدت‌ها محل توجه بسیاری از مقالات شبه‌آزمایشی بوده است (مثلاً Campbell & Stanley, 1963)، اما در مدلسازی سوگیری انتخاب تا همین اواخر مغفول ملنده بود (مثلاً Heckman & Roselius, 1994؛ Heckman et al., 1997). به عنوان مثال، فریدلندر و رابینز (۱۹۹۵) دریافتند که مدل‌های سوگیری انتخاب آزمایش‌های مرتبط با رفاه، زمانی بیشترین شباهت را با برآوردهای آزمایش‌های تصادفی دارند که گروههای کنترل‌های غیرتصادفی از همان ایالتهایی انتخاب شوند که شرکت‌کنندگان برنامه به آن تعلق دارند. بل و همکارانش (Bell et al., 1995) گروه‌های کنترل داخلی مختلفی را مورد بررسی قرار دادند. این گروهها از میان متقاضیانی انتخاب شده بودند که یا حذف شده بودند، یا در غربالگری

[۳۴۲] دهجیا و واهابا (Dehejia and wahba 1999)، داده‌های لالوند (Lalonde 1986) را با استفاده از تحلیل نمره تمایل دوباره بررسی کردند و براوردهای نقطه‌ای بدست آوردند که به براوردهای آزمایش تصادفی معیار بسیار نزدیک‌تر بود.

کنار گذاشته شده بودند، و یا در زمان اجرای مداخله حاضر نشده بودند. مطالعه آنها نتایج دلگرم‌کننده‌ای بدست آورد.

همچنین مدل‌های انتخاب در صورتی عملکرد بهتری خواهند داشت که پیش‌بین‌هایی انتخاب شده برای این مدلها منعکس‌کننده نظریه‌ها و تحقیقات مرتبط با متغیرهای مبنای تخصیص مداخله باشد، رویه‌ای که نیازمند مطالعه ماهیت نفس پدیده‌ی سوگیری انتخاب است (مثلاً Anderman et al., 1995). به عنوان مثال، رینولدز و تمپل (Reynolds & Temple 1995)، با استفاده از مدل‌های سوگیری انتخاب برآوردهایی از اندازه اثر بدست آوردند که کاملاً شبیه برآوردهای بدست‌آمده از آزمایش‌های تصادفی صورت گرفته بر روی اثرات حضور در برنامه پیش‌دبستانی بود. مقررات مربوط به واجد شرایط بودن برای شرکت در این برنامه کاملاً مشخص بود و محققین می‌توانستند مشارکت را به طور دقیق پیش‌بینی کنند. با این وجود، زمانی که اطلاعات کمتری راجع به مشارکت وجود داشت، حتی محققینی که تلاش قابل‌توجهی برای انتخاب گروه کنترل قابل‌قیاس و اندازه‌گیری پیش‌بین‌های انتخاب متناسب داشتند، نتایجی بدست آوردند که این سؤال را در آنها برمی‌انگیخت که آیا مدل سوگیری انتخاب اصلاً درست کار می‌کند یا نه؟ (Grossman & Tierney, 1993).

هکمان (Heckman & Todd, 1996؛ Heckman & Roselius, 1994, 1995؛ Heckman, Lalonde & Smith, 1999) درسهای فراگرفته از مطالعات پیشین را در مدل‌های بازبینی‌شده مختلفی که اثرات اشتغال و برنامه‌های آموزشی را -تحت قانون شراکت در آموزش شغلی[343] (JTPA)- آزمایش می‌کردند، اعمال کردند. این قانون، در واقع آزمایشی ملّی‌ست که در سال ۱۹۸۶ توسط وزارت کار آمریکا تصویب شده است. این آزمایش، شامل شرکت‌کنندگان برنامه، یک گروه کنترل تصادفی، و یک گروه مقایسه غیرتصادفی از میان کسانی است که واجد شرایط شرکت در JTPA بودند، اما درخواست نداده بوند. هکمن، چندین متغیر برآوردگر سوگیری انتخاب نیمه‌پارامتری که به چنین پیشفرضهای مستحکمی نیاز نداشتند را بررسی کرده، و دریافت که [این برآوردگرها] نسبت به مدل‌های پارامتری پیشین، عملکرد بهتری دارند[344]. با این وجود، با داشتن پیشفرضهای کمتر، استنباط ضعیف‌تری حاصل می‌شود، و این سؤال سخت که کدامیک از این پیشفرضها مناسب هستند، باید برای هر مطالعه مورد بررسی قرار گیرد. به هر حال، مدل‌هایی که بهترین عملکرد را در مطالعه هکمن و تاد داشتند، با یک نسخه اصلاح شده از نمرات تمایل که در بخش قبل شرح داده شد، انطباق داده شدند. هکمن و تاد (Heckman & Todd, 1996) اذعان می‌کنند که روش‌های انطباق هنگامی بهتر عمل می‌کنند که (۱) اعضای گروه کنترل، از همان بازار کار محلی شرکت‌کنندگان باشند، (۲) به همان پرسشنامه پاسخ دهند، و (۳) داده‌های

343 Job training partnership act

۳۴۴ همه این آزمایش‌ها، با دانستن خروجی آزمایش تصادفی انجام شدند، به نحوی که این تردید به جا مانده است که در شرایطی که محقق نمی‌داند که پاسخ صحیح کدام است (که زمینه احتمالی تقاضانامه هم می‌باشد) این مدل‌ها چه عملکردی دارند.

مربوط به تعیین‌کننده‌های کلیدی مشارکت در برنامه موجود باشد (صفحه ۶۰). شاید این نتایج بیانگر شروع همگرایی میان مقالات آماری، اقتصادی و ادبیات طراحی طرحهای شبه‌آزمایشی برای درک این مطلب باشد که چگونه می‌توان برآوردهای اثر بهتری را از شبه‌آزمایش‌ها بدست آورد.

مانسکی و همکارانش یک شاخص دیگر برای همگرایی ارائه می کنند (مثلاً Manski, 1990؛ Mask & Nagin, 1998؛ Manski, Sandefur, McLanaha & Powers, 1992). این محققین روشهایی ناپارامتریک برای قرار دادن قیدهایی برروی اثرات مداخلهها در شرایط وجود سوگیری انتخاب شناسایی کردند- چیزی شبیه به کاری که در تحلیل حساسیت انجام می شود. در این روش‌ها، پیشفرضهای خشک و اکیدی روی روش‌های پارامتری هکمان نمی‌گذارند. تلاشهای آنها منتج به مجموعه‌ای از قیود تخمین مداخله شد که بسته به پیشفرضهای صورت‌گرفته تغییر می‌کنند. اما برآوردهایی که کمترین پیشفرضها را دارند، توان آماری را فدا می‌کنند، بنابراین ممکن است دامنه قیدها بی‌اندازه وسیع شوند و برآوردهای نقطه‌ای اثرات مداخله تنها داشتن پیشفرضهای قویتر در مورد فرایند مولد انتساب مداخله و خروجیها بدست بیایند. به این معنی که، در حالتی که یک یا چند مدل انتخاب موجّه را بتوان شناسایی کرد. روزنبام (Rosenbaum, 1995b) پیشنهاد می‌کند قیدهای مانسکی مشابه حد تحلیل حساسیت هستند، که در آنها، شاخص کلیدی سوگیری بالقوه در تحلیل حساسیت (Γ) به سمت ∞می‌رود؛ او با این امر موافق است که قیدها محافظه‌کارانه هستند، اما در عین‌حال این قیود حامل اطلاعاتی نیز هستند. کوپاس و لی (Copas & Li, 1997) رابطه بین مدل‌های انتخاب و تحلیل حساسیت را مورد بحث قرار داده و پیشنهاد می‌کنند که مدل‌های انتخاب، چنان‌به پیشفرضها حساس هستند که‌به جای اینکه از آنها برای برآورد پارامترهای یک مداخله منفرد استفاده کرد، باید آنها را با تغییر دادن حساب‌شده پیشفرضها، به عنوان تحلیل حساسیت به کار گرفت. هکمان و همکارانش نیز با این دیدگاه موافق هستند (مثلاً Heckman & Hotz, 1986؛ Winship & Mare, 1992). همه توافق دارند که تحلیل حساسیت در آزمایش‌های غیرتصادفی حیاتی است.

مدلسازی معادله ساختاری متغیر پنهان

کار کارل یورسکوگ و همکارانش بر روی مدلسازی معادله ساختاری (مانند Joreskog & Sorbom, 1988, 1993) و کار مشابه اما کاربرپسندتر پیتر بنتلر (مثلاً Bentler 1993, 1995) منجر به کاربرد گسترده تکنیک‌های موسوم به «مدلسازی علّی» شد. امید می‌رفت که وقتی این تکنیک‌ها بر داده‌های شبه‌آزمایش‌ها اعمال شدند، با انجام تعدیلات بر روی پیش‌بین‌های خروجی‌ای که با دریافت مداخله همبستگی داشتند، و همینطور با تعدیلاتی که برروی مشکلات (ناپایاییهای) اندازه‌گیری در پیش‌بینها صورت می‌دهند، باعث شوند استنباط‌های علّی دقیق‌تر و با صحت بیشتری حاصل شود. اگر این دو هدف برآورده می شد، تخمینی بدون سوگیری از آثرات مداخله بدست می‌آمد. در واقع، تعدیل خطای اندازه‌گیری با استفاده از مدل‌های متغیر مکنون امکان‌پذیر است (فصل ۱۲ را ببینید). برای انجام این کار، به مقیاسهای اندازه‌گیری مشاهده‌شده متعددی از یک سازه نیاز داریم که، در

حقیقت، مورد تحلیل عاملی قرار می‌گیرند تا متغیرهای مکنونی بدون خطای اندازه‌گیری تصادفی حاصل شود (می‌توان برای کاهش هزینه‌ها، اندازه‌گیریهای متعدد روی یک زیرنمونه (بخشی از یک نمونه) انجام داد ؛ Allison & Hauser, 1991). از این متغیرهای مکنون می‌توان برای مدل کردن نتایج مداخله و بهبود برآورد اثرات مداخله استفاده کرد. به عنوان مثال، تحلیل‌های مجدد متعدد روی داده‌های اصلی مطالعه طرح پیش‌دبستانیها (Cicirelli & Associates, 1969) با کمک متغیرهای مکنون صورت گرفت، و همه آن بازتحلیلها برآوردهایی تولید کردند که بهتر از برآورد اثر بدست آمده از تحلیل اولیه بودند (Bentler & Woodward, 1978؛ Magidson, 1977, 1978, 2000؛ Reynolds & Temple, 1995؛ Rindskopf, 1981).

با این وجود، هدف دیگر این مدل‌ها، که تعدیل خروجیها از نظر متغیرهایی که با متغیر تخصیص و نتایج مداخله همبستگی دارند است، مشکل‌آفرین است، زیرا به ندرت در علوم اجتماعی پیش می‌آید که دانش کافی راجع به همه پیش‌بین‌های خروجی همبسته با مداخله وجود داشته باشد. این ناتوانی اجتناب‌ناپذیر در لحاظ کردن همه این پیش‌بین‌ها، سبب بروز سوگیریهای پنهان، و برآورد نادرست اثرات شود. به علاوه، اینکه مدل شامل همه پیش‌بین‌ها باشد کافی نیست؛ مدل باید رابطه آنها با همدیگر و با خروجی (از جمله عبارات غیرخطی و برهم‌کنشی، مدلسازی صحیح روابط مستقیم و روابط با واسطه، و پیامد مستقیم روابط دارای تأخیر) را به درستی پیش‌بینی کند. ریچارد و گالوب (Reichardt & Gollob, 1986) مقدمه‌ای خواندنی در باب این مسائل نوشته‌اند؛ بولن (Bollen, 1989) کمی بیشتر به جزئیات این موضوع پرداخته، و بنتلر و چو (Bentler & Chou, 1988) نکاتی عملی برای به کارگیری مؤثرتر این مدل‌ها ارائه کرده‌اند. در نهایت، همه (حتی خود سازندگان LISREI که ادعا می‌کنند برنامه آنها، فارغ از آزمون اینکه آیا خود رابطه علّیست یا خیر، پارامترهای علّی که طبق فرض صحیح در نظر گرفته می‌شوند را برآورد می‌کند (Joreskog & Sorbom, 1990) بر سر این موضوع توافق دارند که این مدل‌های علّی، تنها به اندازه طراحی داده‌هایی که به آنها داده می شود، می‌توانند خوب باشند (خوبی عملکرد آنها وابسته به کیفیت داده‌های ورودی به آنهاست).

مقالات مربوط به مدلسازی معادله ساختاری، مستقل از مقالات مربوط به مدلسازی سوگیری انتخاب و نمرات تمایل، توسعه بسیاری پیدا کرده‌اند. این گسترش تا اندازه‌ای به دلیل ادغام رشته‌های مختلف بوده است، و تا حدی به این علت است که در این روش‌ها از تنظیمات متفاوتی بهره گرفته می‌شود: مدل‌های معادله ساختاری، برای پیش‌بین‌های خروجی تنظیم شده‌اند، اما مدل‌های سوگیری انتخاب و نمرات تمایل، برای پیش‌بین‌های انتخاب مداخله تنظیم شده‌اند. با این حال، وینشیپ و مورگان (Winship & Morgan, 1999) به وضوح بیان می‌کنند که رابطه‌ی نزدیکی میان همه این روش‌ها وجود دارد (به مقالات Pearl, 2000؛ Spirtes, Glymour & Scheines, 200 نیز مراجعه کنید). مشخص نیست که آیا این تلاش‌ها برای تلفیق و یکپارچه کردن این عناوین، بدون انجام پیشفرضهایی که تا کنون به عنوان مانعی برای انجام تحلیل‌های به حساب می‌آمده، موفقیت‌آمیز

خواهد بود یا نه. اما همچنان مشخص است که توجه به استنباط علّی، در انواع علوم رو به گسترش است، که این حداقل حکایت از یکپارچه‌گی بیشتر حوزه های واگرای تحقیقاتی دارد.

فصل شش

شبه‌آزمایشها: طرح های سری زمانی قطع شده

زمان (اسم): انگلیسی میانه از ریشه انگلیسی قدیم کلمه tima. ۱. به طور مخفف با حرف T نشان داده می شود. الف. طیفی غیرفضایی که در آن، رویدادها با ترتیبی برگشت‌ناپذیر از گذشته به حال و آینده به وقوع می پیوندند. ب. بازه‌ای که دو نقطه واقع شده در یک طیف را از یکدیگر جدا می‌کند.

زمان (صفت): ۱. مرتبط با محاسبه زمان

سریها: از ریشه کلمه لاتین serere به معنای پیوستن. ۱. تعدادی از اشیاء یا رویدادها که به صورت یکی پس از دیگری مرتب شده، رخ می‌دهند.

در جولای سال ۱۹۸۲، دادگاهی ایالتی آریزونا مجازاتهای بسیار سنگین را برای رانندگی در حال مستی و خلصه ناشی از مصرف مواد مخدر تصویب کرد. مقایسه نتایج ماهانه از ژانویه ۱۹۷۶ تا ژوئن ۱۹۸۲ (شرایط کنترل) با نتایج ماهانه مجموعها در فاصله جولای ۱۹۸۲ تا می ۱۹۸۴ (شرایط مداخله) نشان‌دهنده کاهش مرگ‌ومیرهای ناشی از تصادفات پس از تصویب قانون بود. یافته‌های مشابهی در سن‌دیگو در کالفرنیا بدست آمد، پس از اینکه این شهر در ژانویه سال ۱۹۸۲ مجازاتهایی قانونی را برای رانندگان مست تعیین کرد. در الپاسو در تگزاس (که چنین مجازاتهای را قرار نداده بودند)، آمار ماهیانه مرگ‌ومیر تغییرات مشابهی را در طول ماههای ژانویه تا جولای سال ۱۹۸۲ نشان نمی‌دادند. تغییرات در روندهای مرگ و میر در آریزونا و سن‌دیگو در قیاس با عدم وجود تغییرات در الپاسو نشان می‌دهد که قوانین جدید، مرگ‌ومیر را کاهش داده است (West, Hepworth, McCall, & Reich, 1989). این نوع طرحهای سری‌زمانی قطع‌شده یکی از اثرگذارترین و قدرتمندترین طرحهای شبه‌آزمایشی هستند، علی‌الخصوص زمانی که برخی عناصر کمکی مورد بحث در فصول قبلی به آنها اضافه می‌شود (نگاه کنید به جدول ۵.۲).

سری‌زمانی[345] چیست؟

[345] Time series

سری‌زمانی، به سری بزرگی از مشاهدات انجام‌شده روی یک متغیر به صورت پیوسته در طول یک دوره زمانی گفته می‌شود. ممکن است مشاهدات روی واحدهای یکسانی باشند، مثلاً در حالتی که نشانگان پزشکی یا روانپزشکی یک فرد به صورت مکرر در طول زمان مشاهده می‌شود. یا ممکن است مشاهدات روی واحدهای متفاوت اما مشابه باشند، مثلاً زمانی که تعداد کشته‌شدگان ناشی از تصادف در یک منطقه خاص در طول چند سال بررسی می‌شود (با توجه به اینکه در طی این زمان، جمعیت پایه به صورت پیوسته در حال تغییر بوده است). در این فصل، نوعی خاص از سریهای‌زمانی به نام سری‌زمانی «قطع‌شده[۳۴۶]» را توصیف می‌کنیم که می‌توان برای ارزیابی اثر مداخله از آن بهره گرفت. نکته کلیدی در اینجا، شناختن یک نقطه خاص از سری است که مداخله در آن رخ می‌دهد، مانند روزی که قانون بستن کمربند ماشین اجباری شد. اگر مداخله تأثیر داشته باشد، فرضیه علّی این است که مشاهدات بعد از مداخله، سطح یا شیب متفاوتی نسبت به مشاهدات قبل از مداخله خواهند داشت. یعنی در زمانی که مداخله صورت می‌گیرد، سری باید نسبت به موقعیت پیشین یک انقطاع را نشان دهد. چنین طرح‌هایی، کاربرد گسترده‌ای در ارزیابی اثر مداخلهها در حوزه‌های مختلف دارند؛ که از جمله آنها می‌توان به مواردی همچون تبلیغات وکالت (Johnson, Yazdi & Gelb, 1993)، مداخلههای جامعه-محور برای بهبود فرزندپروری (Biglan, Metzler & Ary, 1994)، اپیدمیولوژی یا همه‌گیرشناسی[۳۴۷] (Catalano & Serxner, 1987؛ Tesoriero, Sorin, Burrows & LaChance-McCullough, 1995)، ایمنی محصول مشتری‌محور (Orwin, 1984)، کنترل سلاح (Carrington & Moyer, 1994؛ O'Carroll et al., 1991)، تاریخچه ازدواج (Denton, 1994)، حقوق بشر (W. Stanley, 1987)، مشارکت سیاسی (Seamon & Feiock, 1995)، ارزش املاک و مستغلات (Brunette, 1995؛ Murdoch, Singh & Thayer, 1993)، تحلیل ریسک محیطی (Teague, Bernardo & Mapp, 1995)، سوء رفتار با همسر (Tilden & Shepherd, 1987)، جراحی (Everitt, Sourmerai, Avorn, Klapholz & Wessels, 1990)، سوء مصرف مواد (Velicer, 1994)، سیاست مالیاتی (Bloom & Ladd, 1982)، ایمنی مکان کار (Feinauer & Havlovic, 1993) و تأثیر تلویزیون (Hennigan et al., 1982)، اشاره داشت. در شرایطی که طرح‌های تصادفی توجیه‌پذیر نبوده، و امکان وجود یک سری‌زمانی محتمل باشد، سری‌زمانی قطع‌شده، جایگزین شبه‌آزمایشی قوی برای طرح‌های تصادفی خواهد بود. در اینجا، با ساده‌ترین طرح سری‌زمانی قطع‌شده کار را شروع کرده، و سپس، نسخه‌های مختلف این طرح را معرفی می نماییم، و در نهایت، مشکلات عملی که در حین اجرای مطالعات سری‌زمانی رخ می‌دهند را مورد بحث و بررسی قرار خواهیم داد.

توصیف انواع آثار

[346] Interrupted
[347] Epidemiology

سری‌زمانی پس از مداخله ممکن است از چندین نظر با سری‌زمانی قبل از مداخله متفاوت باشد. اول اینکه، ممکن است یک ناپیوستگی تند در نقطه اجرای مداخله وجود داشته باشد که در واقع، همان نقطه‌ای است که انتظار داریم سری‌زمانی کلی در آن قطع شده باشد. یک سری‌زمانی کوتاه با ۲۰ مشاهده را در نظر بگیرید که ۱۱ مشاهده قبل از مداخله، و ۹ مشاهده بعد از آن صورت گرفته است. اگر مقادیر سری پیش از مداخله، ۲، ۳، ۴، ۵، ۶، ۷، ۸، ۹، ۱۰، ۱۱ و ۱۲ باشند و مقادیر پس از مداخله، ۱۰، ۱۱، ۱۲، ۱۳، ۱۴، ۱۵، ۱۶، ۱۷ و ۱۸ باشند، آنگاه می‌توان نتیجه گرفت که مقادیر سری بعد از مداخله کاهش یافته‌اند، و مقدار مشاهده ۱۲ام به جای اینکه ۱۳ باشد، ۱۰ است. تغییر مقدار سری از ۱۲ به ۱۰، تغییر در سطح یا عرض[۳۴۸] نامیده می‌شود، زیرا (۱) سطح سری افت می‌کند، و (۲) شیب‌های پیش از مداخله و پس از مداخله، عرضهای متفاوتی دارند. دوم اینکه، ممکن است در نقطه انقطاع، تغییری در شیب سری وجود داشته باشد. فرض کنید مقادیر قبل از مداخله، عبارتند از ۲، ۳، ۴، ۵، ۶، ۷، ۸، ۹، ۱۰، ۱۱ و ۱۲ و مقادیر پس از مداخله عبارتند از ۱۴، ۱۶، ۱۸، ۲۰، ۲۲، ۲۴، ۲۶ و ۲۸. بنابراین، قبل از مداخله، سری در هر بازه زمانی، یک واحد جابه‌جا می‌شده اما بعد از مداخله، در هر بازه زمانی، دو واحد تغییر کرده است. این حالت نام‌های مختلفی دارد و تغییر در سوق[۳۴۹]، روند[۳۵۰] یا شیب[۳۵۱] نامیده می‌شود.

اگرچه تغییر در سطح و شیب، متداول‌ترین شکل‌های تغییر هستند، اما تنها حالت‌های ممکن نیستند. به عنوان مثال، اگر مداخله سبب شود که مردم نسبت به دوره زمانی پیش از مداخله، از یک نظر همگن‌تر یا ناهمگن‌تر شوند، تغییرات پس از مداخله به صورت واریانس‌هایی حول هر میانگین خواهند بود. یا ممکن است یک الگوی دوره‌ای تحت تأثیر قرار گیرد؛ مثلاً معرفی سیستم‌های تهویه هوا به بازار احتمالاً رابطه میان ایام سال و اوقاتی که در داخل منزل گذرانده می‌شود را تغییر می‌دهد. اگرچه محققان معمولاً در تحقیقات سری‌زمانی قطع‌شده، به دنبال تغییرات سطح و شیب هستند، اما باید احتمال وجود آثار دیگر را نیز در نظر داشته باشند.

اثرات را می‌توان در طول دیگر ابعاد نیز مشخصه‌یابی کرد. اثر پیوسته، در طول زمان کاهش نمی‌یابد. بنابراین، همان جابه‌جایی در سطح X واحد که بلافاصله بعد از مداخله بدست‌آمده، همچنان در بخش‌های انتهایی سری‌زمانی حضور دارد. اثر ناپیوسته، در طول زمان پایدار نیست. معمولاً وقتی اثری کاهش پیدا می‌کند، اثر اولیه در طول زمان به سمت سطح یا شیب پیش از اختلال سوق پیدا می‌کند. چه زمانی که مداخله اعمال‌شده و سپس حذف شود، و چه زمانی که یک مداخله با آثار گذرا در آن مکان باقی گذاشته شود، امکان وقوع این حالت وجود دارد. گاهی ممکن است اثر پیوسته حالتی معکوس گرفته، و این اثر در طول زمان بزرگتر شده و یک اثر شدیدتر ایجاد کند (Cook et al., 1979). البته تجربه نشان می‌دهد که این حالت نادر است.

[348] Intercept
[349] Drift
[350] Trend
[351] Slope

اثرات ممکن است بر اساس ماهیت اولیه‌شان، بعد از اجرای مداخله به صورت فوری، یا تأخیریافته مشاهده شوند. معمولاً تفسیر آثار فوری ساده‌تر است، زیرا شروع آنها دقیقاً با زمان مداخله، مطابقت دارد. تفسیر اثرات تأخیریافته سخت‌تر است، مگر اینکه نوعی تبیین نظری وجود داشته باشد که مدت تأخیر قبل از وقوع اثر را بتوان به کمک آن محاسبه کرد (مثلاً زیست‌شناسی به ما می‌گوید که باید بین کاربرد یک روش کنترل بارداری و اولین نتیجه آن اثر، نه ماه صبر کرد). در مورد آثار با تأخیر، هر چه فاصله زمانی میان مداخله و اولین نشانه‌های مشهود تأثیر آن، بیشتر باشد، تفسیرهای جایگزینی که برای آن ارائه می‌شود، متنوع‌تر خواهند بود. بنابراین، اثر سری‌زمانی قطع‌شده باید در سه بعد به صورت همزمان توصیف شود، که این سه بعد عبارتند از: شکل اثر (سطح، شیب، واریانس و چرخه‌ای بودن)، دوام آن (پیوسته یا ناپیوسته) و فوریت آن (فوری یا باتأخیر). ما در این فصل، مثال‌هایی از تمامی این آثار ارائه می‌کنیم.

نکات مختصری در باب تحلیل

ما به جای تحلیل، تقریباً به صورت کامل روی طراحی متمرکز شده‌ایم، اگرچه گاهی برخی از مشکلات در اعتبار نتایج آماری را از طریق مثال‌های خاص نشان می‌دهیم.[۳۵۲] با این وجود، بهتر است که چند نکته پایه را راجع به آمار سری‌زمانی یاد بگیریم. به عنوان مثال، برای مقایسه مشاهدات پیش از مداخله با مشاهدات پس از مداخله با استفاده از آزمون t، نمی‌توان از آمار معمولی استفاده کرد. اما داده‌های سری‌زمانی معمولاً به صورت خودکار تصحیح می‌شوند. یعنی مقدار یک مشاهده می‌تواند با مقدار مشاهده پیشینی که یک، دو، سه یا چندین گام پیشتر رخ داده است، ارتباط داشته باشد. برآورد این تصحیح خودکار، نیازمند مشاهدات زیادی است (معمولاً ۱۰۰) تا

---

[۳۵۲] روش‌های آماری برای تحلیل سری زمانی را می‌توان به دامنه زمانی و دامنه فرکانس تقسیم‌بندی کرد (Shumway, 1988). رویکردهای دامنه زمانی، با پیش‌بینی مقادیر کنونی از مقادیر گذشته، مشاهدات سری زمانی را مدلسازی می‌کنند. بهترین روش شناخته‌شده بر این اساس، رویکرد میانگین متحرک انتگالی اتورگرسیو (ARIMA) است که توسط باکس و جنکینز (Box and Jenkins, 1970) ارائه شده است. تحلیل‌های اقتصادی متعددی بر این اساس ارائه شده‌اند که گاهی آنها را مدل‌های رگسیون ساختاری می‌نامند (Kim and Trivedi, 1994؛ Ostrom, 1990). رویکردهای دامنه فرکانس، مشاهدات را به صورت ترکیبی از امواج سینوسی و کسینوسی متناوب مدلسازی می‌کنند که گاهی تحلیل طیفی، هماهنگ یا فوریه هم نامیده می‌شود، و این روش در علوم و مهندسی فیزیک بسیار پر کاربرد است. شاموی (Shumway, 1988) ادعا کرد که اگر سری زمانی طولانی باشد، هر دو رویکرد زمانی و فرکانسی نتایج مشابهی حاصل می‌کنند. در کل، اجماعی برای ترجیح یک روش نسبت به روش دیگر وجود ندارد. تحلیل‌گران به سرعت در حال توسعه سری‌های زمانی چندمتغیره، بررسی داده‌های از دست رفته، سری زمانی غیرخطی، سری زمانی تجمیع شده که ترکیبی از سری‌های زمانی متعدد در واحدهای مختلف است و براورد سری زمانی که نسبت به نقاط دورافتاده استوار است، می‌باشند (Box, Jenkins & Reinsel, 1994؛ Caines, 1988؛ Hannan & Deiseler, 1988؛ Kendall & Ord, 1990؛ Sayrs, 1989؛ Tong, 1990؛ W. Wei, 1990). مجله تحلیل سری زمانی پیشرفت‌های نوین این رشته را ااائه می‌کند و مجله انجمن آمار آمریکا، مقالاتی را به صورت منظم در این باره منتشر می‌کند. کتاب‌های سری زمانی در سطح پایه (Cromwell, Labys & Terraza, 1994؛ Cromwell, Hannan, labys & Terraza, 1994؛ McDowell, McCleary, Meidinger & Hay, 1980؛ OStrom, 1990؛ Sayrs, 1989) و پیشرفته (مانند Box, Jenkins & Reinsel, 1994؛ Fuller, 1995؛ Hmailton, 1994؛ Harvey, 1990؛ Judge, Hill, Griffiths & Lee, 1985؛ Kendall & Ord, 1990؛ Reinsel, 1993؛ W. Wei, 1990) ارائه شده‌اند. برخی از آنها حاوی برنامه‌های نرم‌افزار رایانه‌ای هستند (مانند Brockwell & Davis, 1991؛ Shumway, 1988) و بسته‌های آماری استاندارد، توانایی تحلیل سری زمانی را توسعه داده‌اند (Harrop & Velicer, 1990a, 1990b؛ Newbold, Agiakloglou & Miller, 1994؛ Kim & Trivedi, 1994).

بتوان به راحتی مدل صحیح را شناسایی کرد (Velicer & Hartop, 1983 ؛Box, Jenkins & Reinsel, 1994). با این وجود، تعداد دقیق مشاهدات لازم را نمی‌توان از قبل تعیین کرد. از نظر آماری، گاهی فقط به داده کافی برای شناسایی مدل نیاز داریم، و این شناسایی به میزان خطای داده‌ها، هر گونه آثار تناوبی، زمانبندی مداخلات درون سری، و تعداد تأخیرهایی که باید مدلسازی شوند، بستگی دارد. اگرچه معمولاً در مقالات، وجود ۱۰۰ مشاهده را مطلوب در نظر می‌گیرند، ما در این فصل مثال‌هایی ارائه می‌کنیم که تعداد مشاهدات کمتری دارند[۳۵۳]. دلیلش این است که گاهی می‌توان مدل را با تعداد مشاهدات کمتر شناسایی کرد. با این وجود، معمولاً از سری‌زمانی کوتاه به این منظور بهره می‌گیریم که نشان دهیم در مواردی که تنها چند نقطه زمانی پیش یا پس از آزمون وجود دارد، چگونه یک سری زمانی کوتاه‌شده می‌تواند مانع از در خطر افتادن بیشتر اعتبار محاسبات شود. تفسیر سری‌زمانی منحصراً یک امر آماری نیست؛ بلکه به ویژگی‌های طراحی و تأخر[۳۵۴] و بزرگی اثر هم بستگی دارد.

سری زمانی قطع‌شده ساده

یک طرح پایه برای سری زمانی، نیازمند یک گروه مداخله است که در آن، مشاهدات زیادی قبل و بعد از مداخله انجام شده است. طرحی با ده مشاهده را می‌توان به صورت زیر رسم کرد:

O_1	O_2	O_3	O_4	O_5	X	O_6	O_7	O_8	O_9	O_{10}

تغییر در سطح

اولین مثالی که از سری‌زمانی قطع‌شده ساده ارائه می‌کنیم (McSweeny, 1980)، ۱۸۰ مشاهده دارد، و برای کنار گذاشتن مقادیری که اعتبار درونی را تهدید می‌کنند، یک حالت ایده‌آل است (شکل ۶.۱). در مارس ۱۹۷۴، شرکت سینسیناتی‌بل[۳۵۵]، برای هر بار تلفن به راهنمای تلفن محلی، هزینه ۲۰ سنت در نظر گرفت. شکل ۶.۱، افت شدید و بلافاصله در تماس‌های گرفته شده با راهنمای تلفن محلی را بعد از اعمال این هزینه نشان می‌دهد. با اندکی تأمل می‌توان چند فرضیه جایگزین برای این قضیه ارائه کرد. تهدید رگرسیون به میانگین به نظر توجیه منطقی ندارد، زیرا سری‌های زمانی بسیار طولانی پیش از مداخله نشان می‌دهد که تعداد تماسهای زیاد با راهنمای تلفن در طول چندین سال بوده و نه بلافاصله قبل از مداخله (برای مشاهده سری‌زمانی با مصنوعات[۳۵۶] رگرسیونی بالا به Maltzz, Gordon, McDowell & McCleary, 1980 مراجعه کنید). اگر جمعیتی که با راهنمای تلفن محلی در سینسیناتی تماس گرفته‌اند، در ماه‌های قبل و بعد از مداخله تغییری نداشته و نمونه‌های قبل و بعد از مداخله

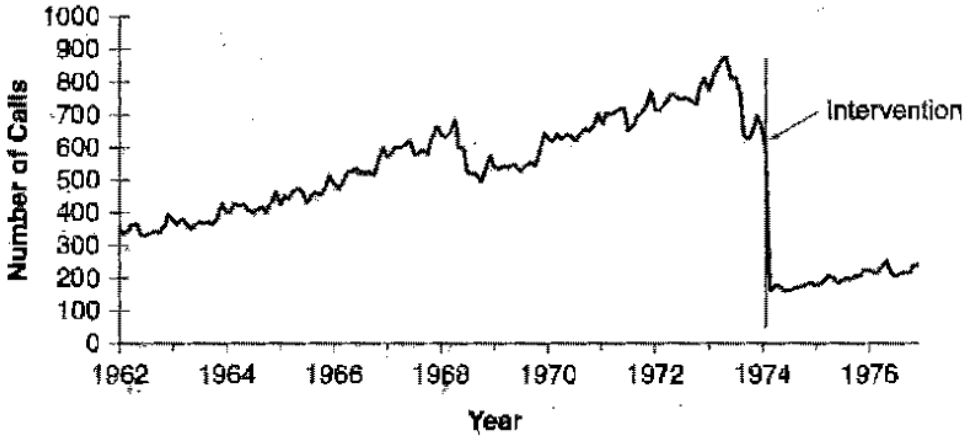
[۳۵۳] غیر از حالاتی که به صورت صریح بیان شده باشد، هیچ کدام از شکل‌های این فصل داده‌های خام در نقاط زمانی را تجمیع نکرده‌اند؛ یعنی همه نقاط زمانی موجود در داده‌های خام در هر شکل وارد شده‌اند.

[354] Immediacy

[355] Cincinati Bell

[356] Artifacts

برای هر کدام از متغیرهای مؤثر بر نتایج (خروجی)، یکسان باشند، سوگیری انتخاب نیز غیرموجّه خواهد بود. همچنین خطر ریزش[۳۵۷] نامحتمل است، زیرا به نظر نمی‌رسد که تعداد زیادی از مشتریان به خاطر این هزینه، تلفن خود را قطع کنند، و این امر را می‌توان به آسانی با کمک سوابق شرکت بررسی کرد. همچنین هیچ فرایند بلوغ طبیعی شناخته‌شده‌ای نمی‌تواند باعث چنین افت شدیدی در استفاده از راهنمای تلفن محلی شود. اثرات آزمون[۳۵۸] نیز نامحتمل به نظر می‌رسند، چون در اینجا، آزمایش بر اساس صورتحساب است و اثر آزمون مستلزم این است که شرکت تلفن نحوه ارائه صورتحساب را تغییر دهد، به گونه‌ای که تعداد تماس‌های گرفته شده با راهنمای تلفن قبل از اعمال هزینه بیان شود، و مشتری بر اساس بازخوردی که از آن می‌گیرد (نه بر اساس هزینه)، رفتار خود را تغییر دهد. می‌توان بررسی کرد که آیا این کار انجام شده است یا خیر. سوگیری ناشی از گذشت زمان تنها زمانی موجّه خواهد بود که بتوان رویداد دیگری را پیدا کرد که همزمان با اعمال هزینه ۲۰ سنتی رخ داده، و می‌توانسته چنین اثر بزرگی داشته باشد، چیزی که به نظر نامحتمل می‌رسد. وقتی که آثار اینگونه بلافاصله و شدید هستند، اغلب تهدیدهایی[۳۵۹] روایی درونی نامحتمل می‌شوند. مثال‌های دیگری از سری‌زمانی قطع‌شده با آثار بلافاصله و شدید، عبارتند از اثر واکسیناسیون در بیماری‌هایی مانند کزاز (Veney, 1993) و غربالگری فنیل کتونوری (PKU) بر عقب‌ماندگی (MacCready, 1974).



شکل(۶.۱). آثار تعیین هزینه بر راهنمای تلفن در شرکت سینسیناتی‌بل.

تغییر در شیب

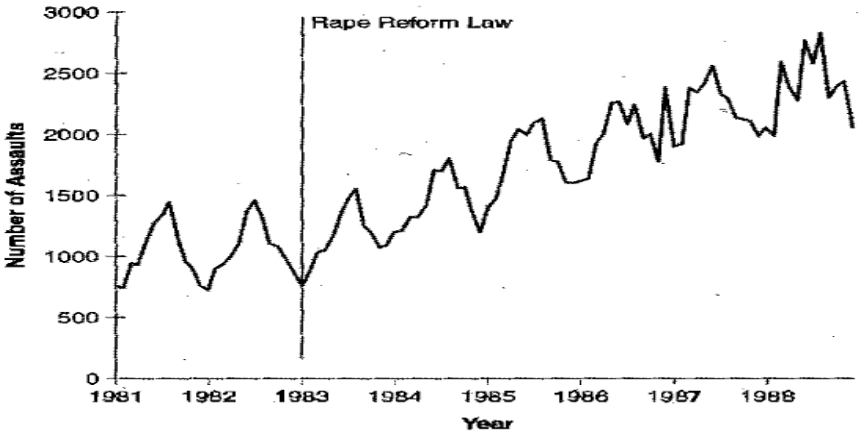
[357] Attrition
[358] Testing effect
[359] Threat

شکل ۶.۲، یک سری‌زمانی قطع‌شده را با ۹۶ مشاهده نشان می‌دهد که خروجی آن، به جای تغییر سطح، تغییر شیب است. در سال ۱۹۸۳، کانادا قانون جنایی مربوط به تعرض را اصلاح کرد. قانون قدیمی به دو دسته تقسیم-بندی شده بود: تجاوز و تعرض. قانون جدید شامل سه گروه به ترتیب افزایش شدت بود، و تهمیدات دیگری نیز برای افزایش تعداد قربانیانی که جنایت را به پلیس گزارش می‌کنند، اندیشیده شده بود و با عمومیت گسترده‌تری اجرا می‌شد. رابرتز و گبوتیس (Roberts and Gebotys, 1992) برای ارزیابی تأثیر این تغییر، از داده‌های ماهانه سیستم گزارش جرم سراسری کانادا برای تعرض در دوره ۱۹۸۱ تا ۱۹۸۸ بهره گرفتند. سری زمانی حاصله، یک اثر فصلی را نشان می‌داد، بدین صورت که تعرضات بیشتری در تابستان و تعرضات کمتری در زمستان گزارش می‌شد. باید توجه داشت که قبل از تصویب این قانون، شیب نزدیک به صفر بود. بعد از تصویب قانون، شیب سری‌زمانی افزایش یافت که نشان می‌داد این قانون، آثار مطلوبی روی گزارش تعرضات داشته است.

با توجه به الگوی نتایج، اغلب تهدیدهای روایی، ناموجّه به نظر می‌رسیدند. به عنوان مثال، بلوغ نامحتمل بود زیرا شیب در نقطه مداخله به شدت تغییر کرده بود، و این نتیجه مشابه فرایندهای بلوغ شناخته‌شده نبود. برای اینکه سوگیری گذشت زمان محتمل باشد، یک رویداد دیگر که اثر شدیدی روی گزارش تعارضات داشته، باید وجود داشته باشد. به استثنای عمومیت اجرای قانون (که بخش قابل‌بحث این مداخله است)، هیچ رویداد دیگری از این دست را نمی‌توان شناسایی کرد. با این وجود، نویسندگان داده‌هایی را ارائه کرده‌اند که نشان می‌دهد عمومیت اجرای قانون می‌تواند بر نگرش مردم نسبت به تعرض تأثیر داشته باشد، و می‌تواند در افزایش تعداد گزارش‌ها نقش داشته باشد. این افزایش، روایی سازه روش عملیاتی را مورد سؤال قرار می‌دهد. این مداخله را باید با عبارت «قانون‌گزاری اصلاحی» تعبیر کرد یا با عبارت «قانون‌گزاری اصلاحی با عمومیت بالا»؟ نظر نویسندگان مورد دوم است.

مداخله گروه‌های گزارش‌شده را تغییر داد، بنابراین رابرت و گبوتیز (Roberts and Gebotys, 1992)، چهار تهدید ابزار احتمالی را شناسایی کردند. اول، اینکه قانون جدید به زنان اجازه می‌داد که شوهرانشان را به تعرض متهم کنند، و این اتهام هم می‌توانست بر علیه مردان و هم بر علیه زنان باشد. با در نظر گرفتن این مسئله، نویسندگان مشاهده‌کردند که تعداد پرونده‌هایی که در آنها، مظنون یک زن یا شوهر است، بعد از این قانون تنها ۵ درصد افزایش یافته، که نمی‌تواند افزایش بالای مشاهده شده در شکل ۶.۲ را توجیه کند. دوم این احتمال وجود دارد که جرم‌هایی که پیشتر تحت عنوان «تعرضات دیگر» دسته‌بندی می‌شدند (و در نتیجه در بخش پیش از مداخله شکل ۶.۲ وارد نشده‌اند) اکنون به این گروه تعرض اضافه شده باشند، و تنها در مجموعه پس از مداخله خود را نشان داده باشند، که برای اولین بار، بهره‌برداری و وحشیگری نیز به آن افزوده شده است. اما تحلیل‌هایی که در طول زمان صورت گرفتند هیچ تغییری را در گزارشات مربوط به این دسته جرم‌های ویژه نشان ندادند، و بنابراین احتمال کمی وجود داشت که این دسته از دلایل علّت افزایش مشاهده‌شده در شیب باشند. سوم، شاید این

گزارش‌های مربوط به دیگر انواع تعرضات قبل از اصلاح به ندرت گزارش می‌شدند، اما بعد از اصلاحات، به علّت عمومیت قانون جدید، بیشتر گزارش می‌شوند. نویسندگان نتوانستند هیچ داده‌ای پیدا کنند که بر اساس آن، این تهدید را به صورت مستقیم حذف کنند، اما گزارش کردند که داده‌های غیرمستقیم بیانگر این است که احتمال این تهدید اندک است. چهارم، برخی از داده‌ها بیانگر افزایش موقت در تعداد تعرضات علیه کودکان و نوجوانان بودند. آیا افزایش نشان داده شده در شکل ۶.۲، به بزه‌های علیه نوجوانان محدود می‌شود؟ متأسفانه، آمار ملی را نمی‌توان بر اساس سن تفکیک کرد تا این احتمال آزمایش شود. اما تفکیک آمار جرایم در شهرهایی مانند مونترال نشان می‌داد که افزایش تعداد تعرضات علیه نوجوانان خیلی کمتر از آن است که بتوان آن را دلیل افزایش بالای تعرضات گزارش شده در شکل ۶.۲ دانست.



شکل ۶ – ۲. آثار اصلاح قانون تعرضات در کانادا

“Reforming rape laws: Effects of legislative change in Canada,” by V Roberts and R. J. Gebotya, 1992, Law and Human Behavior, 16, 555-573. Copyright 1992 by Kluwer Academic/Plenum Publishers.

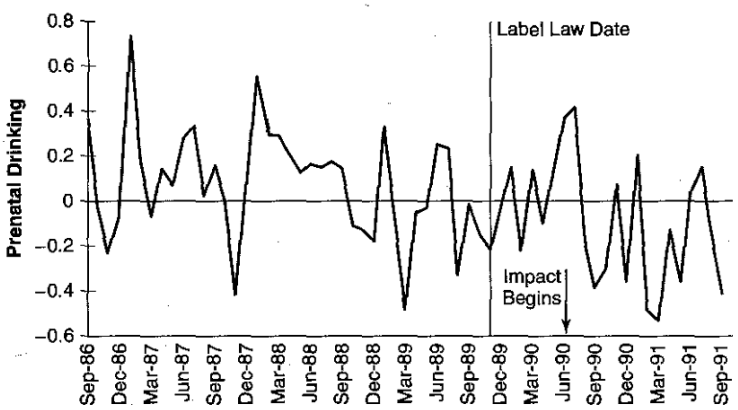
آثار ضعیف و با تأخیر

اگرچه تفسیر شکل ۶.۲ به اندازه شکل ۶.۱ روشن نیست، اما چند طرح سری‌زمانی قطع‌شده ساده را می‌توان در آن مشاهده کرد. شکل ۶.۳، یک مثال رایج‌تر را نشان می‌دهد. این مثال، شامل ۶۳ مشاهده است و به نظر می‌رسد که آثار هم دچار تأخیر در شروع هستند، و هم از نظر شدت ضعیف می‌باشند. در این مطالعه (Hankin et al., 1993)، آثار قانون برچسب هشدار روی بطری مشروبات الکلی بر دفعات نوشیدن الکل توسط زنان حامله بررسی شد. قانون فدرالی که از ۱۸ نوامبر ۱۹۸۹ اجرایی شده بود، چسباندن برچسب هشدار روی بطری نوشیدنی الکل را اجباری می‌کرد. برچسب به صورت مشخص هشدار می‌داد که نوشیدن الکل در دوران حاملگی می‌تواند سبب ایجاد نقص در کودکان شود. نویسندگان، میزان مصرف الکل توسط ۱۲۰۲۶ زن آفریقایی‌آمریکایی حامله، در طول دو هفته قبل از اولین ویزیت حاملگی در کلینیک دیترویت را، قبل و بعد از اجرایی شدن قانون بررسی کردند.

سری‌زمانی، از سپتامبر ۱۹۸۶ تا سپتامبر ۱۹۹۱ را به صورت بازه‌های ماهانه پوشش می‌داد. با بررسی بصری سری زمانی، امکان تفسیر قطعی راجع به تأثیر برچسب هشدار بر روی نوشیدن الکل وجود نداشت.

با این وجود، تحلیل آماری نویسندگان بیانگر این بود که این قانون به جای تأثیر فوری، یک تأثیر باتأخیر داشته است، و هفت ماه بعد از اجرایی شدن، اثر آن آغاز شده است. دلیل وجود این تأخیر این است که قانون تنها بر روی بطریهای تازه تولید اجرا شده بود، نه آنهایی که پیشتر در مغازه‌ها بودند. بنابراین، قبل از اینکه بطریهای دارای برچسب هشدار به تعداد کافی در مغازه‌ها قرار گیرند، و توجه مشتریان را جلب کنند، مدت زمانی سپری شده بود. بر اساس این فرضیه، نویسندگان از زنان پرسیدند که آیا آنها از وجود برچسب آگاه بودند، و آن را مشاهده کرده‌اند؟ تا مارس ۱۹۹۰ یعنی چهار ماه بعد از اجرایی شدن قانون، آگاهی آنها افزایش نیافته بود. بنابراین، انتظار می‌رفت که اثر مداخله باتأخیر مشاهده شود، اگرچه لزوماً تاخیر ۷ ماهه تصور نمی‌شد. به علاوه، به نظر می‌رسید که قبل از مداخله هم، نوشیدن الکل توسط زنان حامله کاهش‌یافته، و این بی‌نظمی می‌تواند یک تهدید احتمالی برای روایی یا اعتبار باشد، با توجه به اینکه این اثر، سبب تغییر تدریجی در شیب می‌شود نه تغییر در سطح. با این وجود، سری زمانی این امکان را ایجاد می‌کرد که اندازه شیب قبل از مداخله را با اندازه بعد از مداخله مقایسه کنیم. این مقایسه نشان داد که تأثیر (باتأخیر) قانون، کاهش مصرف الکل که پیشتر در زنان حامله مشاهده شده بود را تسریع کرده است.

روند فصلی را نیز می‌توان در شکل ۶.۳ دید. نرخ مصرف الکل در تعطیلات انتهای سال و در تابستان بالا می‌رود. با این وجود، فصلی بودن نمی‌توانست تهدیدی برای روایی باشد، زیرا با توجه به اینکه اجرایی شدن قانون نزدیک به فصل کریسمس بود، نوشیدن الکل بلافاصله بعد از اعمال قانون، افزایش هم داشت. از آنجا که انتظار می‌رفت تغییر قانون، نوشیدن الکل را کاهش دهد، اما روند فصلی بر خلاف این فرضیه عمل کرد، در نتیجه نمی‌توان روند فصلی را توضیح‌دهنده کاهش مشاهده‌شده دانست. در واقع، شکل ۶.۳ نشان می‌دهد که مصرف الکل در تعطیلات زمستانی و تابستان بعد از اجرای قانون، نسبت به قبل از قانون، بسیار کمتر است. با این وجود، اگر قانون در فوریه و بعد از تعطیلات یا در پایان تابستان در سپتامبر اجرا می‌شد، این آثار فصلی می‌توانست به غلط به عنوان آثار مداخله تفسیر شود. تحلیل مناسب، نیازمند مدلسازی و حذف آثار فصلی از سری زمانی قبل از ارزیابی اثر مداخله است.



شکل ۶ – ۳. اثر برچسب هشدار الکل روی مصرف الکل در حاملگی.

"A time series analysis of the impact of the alcohol warning label on antenatal drinking," by J. R. Hankin et al., 1993, Alcoholism: Clinical and Experimental Research, 17, pp. 284-289. Copyright 1993 by Lippincott, Williams & Wilkins.

تهدید اعتبار های معمول

در اغلب طرح‌های سری‌زمانی ساده، تهدید اصلی روایی درونی، گذشت زمان است (این احتمال که نیروهایی غیر از مداخله تحت بررسی، همزمان با انجام مداخله، بر متغیر نتیجه‌ای مدنظر تأثیر گذاشته باشند). به عنوان مثال، در داده‌های هانکین و همکارانش (Hankin et al., 1993) می‌توان دید که اگر شهر دیترویت، قانون محدودیت فروش الکل را همزمان با این قانون تصویب می‌کرد، کاهش مصرف الکل در میان زنان حامله به صورتی مشابه با اثر برچسب هشدار رخ می‌داد. می‌توان کنترل‌های مختلفی روی گذشت زمان اعمال کرد که شاید بهترین آنها، افزودن یک سری زمانی بدون مداخله از گروه کنترل باشد که در ادامه به صورت مختصر به توضیح آن خواهیم پرداخت. اما این کار همیشه لازم نیست. به عنوان مثال، مقادیر مصرف الکل در مقاله هانکین و همکارانش، به صورت بازه‌های ماهانه بود و رویدادهای تاریخی که اثر تداخلی ظاهری دارند معمولاً به جای بازه ماهانه، به صورت بازه سالانه هستند. همچنین اگر فهرستی از رویدادهای مؤثر احتمالی تأثیرگذار بر افراد به صورت یک شبه‌آزمایش تهیه شود، با یک ابزار کیفی و کمّی می‌توان تعیین کرد که آیا اغلب یا همه آنها، در بازه بین آخرین پیش‌آزمون و اولین پس‌آزمون عملیاتی شده‌اند یا خیر. اگر اینگونه نباشد، گذشت زمان را نمی‌توان به عنوان تهدیدی در نظر گفت.

تهدید دیگر، تهدید ابزار است. به عنوان مثال، گاهی تغییر در رویکردهای مدیریتی منجر به تغییر در نحوه نگهداری داده‌های ثبت شده می‌گردد. اشخاصی که بخواهند عملکرد خود را خوب نشان دهند، روند حسابداری را تغییر می‌دهند تا عملکرد یا رضایت را بازتعریف کنند. یا افرادی که مأموریت تغییر سازمانی دارند، ممکن است تغییر در

نحوه نگهداری داده‌ها یا تعریف ضوابط موفقیت و شکست را جزو وظایف خود بدانند. به عنوان مثال، به نظر می‌رسد زمانی که اورلاندو ویلسون به ریاست پلیس شیکاگو رسید، این مسأله اتفاق افتاد. با بازتعریف نحوه طبقه‌بندی جرایم، اینطور به نظر رسید که وی جرایم را افزایش داده است. اما افزایش واقعی نبود و تنها تغییر در نحوه ثبت را نشان می‌داد نه رفتار جنایی. همچنین، سری‌زمانی مقاله هانکین و همکارانش (۱۹۹۳)، متکی بر اظهارات خود زنان بود. خوداظهاری در معرض اثر «مشخصات موردتقاضا» است، و عمومیت یک قانون نیز ممکن است چنین تقاضایی را تشدید کند. یعنی خوداظهاری زنان راجع به نوشیدن الکل تحت تأثیر قانون کاهش می‌یابد (حتی اگر میزان مصرف الکل آنها در واقع تغییری نکرده باشد) زیرا زنان می‌دانند که نوشیدن الکل در زمان حاملگی از نظر اجتماعی، مطلوب نیست.

در صورتی که ترکیب گروه آزمایشی در زمان مداخله تغییر قابل‌توجهی پیدا کند، سوگیری انتخاب هم می‌تواند به عنوان تهدیدی دیگر قلمداد شود. اگر مداخله سبب انحراف از چارچوب اندازه‌گیری شود (یا مستلزم آن باشد) این امر امکان‌پذیر می‌شود. اگر اینگونه باشد، انقطاع مشاهده‌شده در سری‌زمانی می‌تواند ناشی از متفاوت بودن افراد در زمان پیش و پس از مداخله بوجود آمده باشد. گاهی می‌توان این مشکل را با محدود کردن تحلیل داده به زیرمجموعه‌هایی که در همه دوره‌های زمانی اندازه‌گیری شده‌اند، رفع کرد. اگرچه انجام این کار همواره ممکن نیست (مثلاً زمانی که نمرات کلاس سوم در یک مدرسه در طول ۲۰ سال مدنظر باشد؛ کودکان کلاس سوم معمولاً هر ساله تغییر می‌کنند). از سوی دیگر، مشخصات واحدها را می‌توان بر این اساس تحلیل کرد که آیا همزمان با انجام مداخله، ناپیوستگی تندی در مشخصات (پروفایل) واحدها دیده می‌شود یا نه.

تهدیدهای روایی نتایج آماری از جمله توان پایین، فرضیات آزمون نقض شده، و عدم وجود امکان اندازه‌گیری مجدد را می‌توان همانند سری‌زمانی برای هر طرح دیگری بررسی کرد. اما مطالعه برچسب ضد مصرف الکل در دیترویت یک مشکل خاص دیگر را نیز نشان می‌دهد. تحلیل‌گران سری‌زمانی باید مشخص کنند که انقطاع در چه نقطه‌ای از سری‌زمانی شروع شده است، و باید نظریه‌ای یا داده‌هایی داشته باشند که بوسیله آنها بتوان نحوه یا الگوی انتشار این آثار به واحدهای در معرض مداخله را ترسیم کرد. در مثال دیترویت، مداخله به کندی نفوذ پیدا کرده، و سرعت دقیق و طرح دقیق انتشار نامشخص است. بنابراین، محققان این صلاحیت را دارند که تعیین کنند در چه زمانی مداخله شروع شود، و می‌توانند با انتخاب زمان شروعی که با بیشینه اثر ظاهری منطبق باشد، شانس خود را تا حد زیادی افزایش دهند. از آنجا که برخی از تغییرات در یک سری‌زمانی طولانی به صورت شانسی رخ می‌دهند، اگر نقطه مداخله به دقت تعیین نشود، منطق علّی طرح بسیار تضعیف می‌شود (اگر سرعت انتشار مشخص باشد، می‌توان به جای شروع ناگهانی تغییر، از منحنی انتشار برای مدل کردن مداخله بهره گرفت. در ادامه فصل درباره این گزینه صحبت خواهیم نمود).

کسانی که از سری‌زمانی بهره می‌گیرند، باید در برابر تمامی مشکلات عمومی روایی سازه مانند تبیین ناقص سازه یا اختلاط در سازه، مصون باشند. با این وجود، خود سری زمانی نیز منجر به ایجاد مشکلات روایی خاصی می‌شود. بسیاری از سری‌های زمانی از داده‌های بایگانی‌شده بهره می‌گیرند، مانند داده‌های رانندگی یا نمرات مدرسه. در این موارد، تهدیدهای مرتبط با واکنش‌پذیری کم‌اثرتر هستند، زیرا تأثیرگذاری پاسخگویان بر مقدار خروجی کار دشواری است. در واقع، پاسخگویان معمولاً نمی‌دانند که بخشی از یک مطالعه هستند. واکنش‌پذیری در سری زمانی بالینی محتمل‌تر است، به ویژه اگر بازه زمانی بین مشاهدات کوتاه بوده و پاسخگویان بتوانند پاسخ‌های قبلی خود را به یاد آورند. بنابراین هر آزمایش سری‌زمانی باید بر اساس معیارهای خودش قضاوت شود تا تعیین شود که آیا نتایج مشاهده‌شده ناشی از نحوه ارزیابی هستند یا مشخصات موردتقاضا، یا دیگر تهدیدهای روایی سازه مشابه.

همچنین در مورد روایی سازه، سری‌زمانی تنها از یک مقیاس اندازه‌گیری متغیر خروجی بهره می‌گیرد. این تا حدودی بدلیل مسائل هزینه‌ای ست، همچنین تفکر سنتی و نادرست نسبت به عملیاتی کردن مفهومی می‌تواند باعث شود تا سازمان‌دهندگان آرشیوها مفاهیمی همچون پیشرفت دانشگاهی، سرقت یا بیکاری را به صورتی مشابه اندازه‌گیری کنند. این مسأله به نوبه خود بغرنج‌تر نیز می‌شود، زیرا معمولاً محقق وادار می‌شود که از مقادیرهای خروجی (نتایج) موجود در دسترس استفاده کند، حتی اگر این مقیاسها و اندازه‌گیری‌های مرتبط با آنها ارتباط کاملی با مداخله تحت‌آزمایش نداشته باشند. این امر سبب می‌شود که مقادیر موجود نسبت به حالتی که خود محققان داده‌های متغیر را جمع‌آوری می‌کنند، یا مقیاسها را متناسب با نظریه مداخله تعدیل می‌کنند، یا می‌توانند گویه های بیشتری را به مقیاس اضافه کنند تا پایایی و روایی را افزایش دهند، از حساسیت کمتری برای شناسایی نتایج برخوردار باشند. البته، در آن‌دسته از سری‌های زمانی که در آنها برای یک اثر چندین مقیاس وجود داشته، و هر یک از این مقیاسها به میزان قابل قبولی واجد روایی هستند، می‌توان تغییر در سری‌زمانی را به صورت جداگانه برای هر مقیاس بررسی کرد. به علاوه، مداخلهها معمولاً رویدادهایی هستند که پاسخ‌دهندگان آنها را طبیعی در نظر می‌گیرند، مانند تغییر قانون، و خروجی‌ها معمولاً با زحمت کمتری گردآوری می‌شوند، زیرا پاسخگویان به گردآوری داده از طرف دولت و شرکت‌ها عادت دارند (حداقل نسبت به انواع تحقیقات دیگر). بنابراین تهدید واکنش‌پذیری کمتر می‌تواند روایی سازه مداخله و نتایج را به خطر بیاندازد.

در مورد روایی بیرونی، گاهی می‌توان روایی بیرونی را با داده‌های موجود در دسترس در مورد مشخصات زمینه‌ای افراد یا واحدها مورد بررسی قرار داد؛ بدین صورت که واحدها را مثلاً بر اساس زن و مرد یا گروه‌های سنی تقسیم‌بندی کرد تا مشاهده شود که آیا اثرات با وجود این تغییرات باز هم برقرار هستند یا خیر. نیازی نیست که این بررسی را به متغیرهای شخصی محدود کنیم. مثلاً می‌توان از متغیرهای مربوط به مختصات آزمایشی برای بدست آوردن محدوده یک اثر استفاده کرد. همچنین می‌توان از متغیرهای زمانی بهره گرفت تا مشخص شود که

آیا یک اثر در زمان‌های مختلفی از روز برقرار است یا خیر (دستگیری در طول روز یا شب). این تفکیک را باید با دقت انجام داد زیرا با ایجاد زیرگروه، توان آماری کاهش می‌یابد. به علاوه، در تحقیقات آرشیوی، انعطاف‌پذیری محدودی در ایجاد زیرگروه وجود دارد (متغیرهای لازم و نقاط قطع باید در میان داده‌های ذخیره‌شده وجود داشته باشند). بنابراین اگر آخرین دسته سنی، «بالای ۶۵ سال» است، و محقق علاقمند به مطالعه افراد خیلی پیر (بالای ۷۵ سال) باشد، امکان انجام این کار برای وی وجود ندارد.

افزودن عناصر طراحی اضافی به طرح سری زمانی قطع‌شده پایه

در فصل‌های پیشین، نشان دادیم که چگونه می‌توانیم با افزودن دقیق مؤلفه‌های طراحی انتخابی به طرح‌های شبه‌آزمایشی پایه، استنباطهای علّی قوی‌تری تولید کنیم. همانگونه که در مثال‌های زیر نیز دیده می‌شود، همین اصول در مورد سری‌زمانی قطع‌شده نیز مصداق دارد.

افزودن سری‌زمانی گروه کنترل غیرهم‌ارز بدون‌مداخله

تعریف سری‌زمانی گروه کنترل را برای یک طرح سری‌زمانی قطع‌شده ساده در نظر بگیرید. طرح حاصله در ذیل رسم شده است:

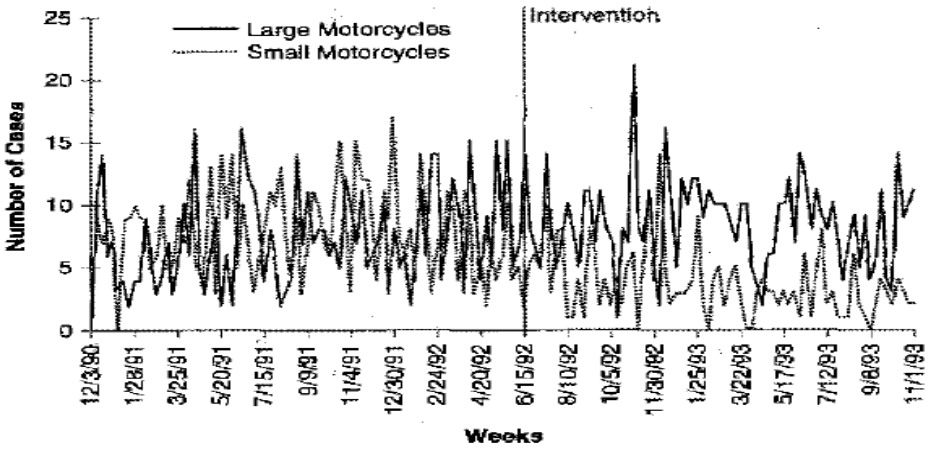
O_1	O_2	O_3	O_4	O_5	X	O_6	O_7	O_8	O_9	O_{10}
O_1	O_2	O_3	O_4	O_5		O_6	O_7	O_8	O_9	O_{10}

مثالی از این طرح در شکل ۶.۴ ارائه شده است. در ژوئن سال ۱۹۹۲، شهر بارسلونا در اسپانیا قانونی را تصویب کرد که طبق آن، رانندگان موتورسیکلت‌های کوچک باید کلاه ایمنی استفاده می‌کردند. استفاده از کلاه ایمنی برای موتورسیکلت‌های بزرگ چندین سال قبل اجباری شده بود. بالارت و ریبا (Ballart and Riba, 1995) از یک سری زمانی با ۱۵۳ مشاهده بهره گرفتند تا تأثیر قانون مذکور را مورد بررسی قرار دهند. متغیرهای نتیجه‌ای (وابسته) طرح، عبارت بود از تعداد قربانیان تصادف موتورسیکلت با آسیب‌های جدی یا مرگ. هر دو متغیر خروجی باید با اجرایی شدن قانون کاهش پیدا کنند، البته این نتیجه باید تنها در مورد قربانیان تصادف با موتورسیکلت‌های کوچک دیده شود. این اثر کاهشی نباید برای گروه کنترل، و رانندگانی که موتور سیکلت بزرگ می‌راندند (که این قانون از چند سال قبل برای آنها اجرا شده بود) مشاهده شود. شکل ۶.۴ نشان می‌دهد که قانون دارای اثر مفروض بوده، و تحلیل آماری از این تفسیر پشتیبانی می‌کند.

از آنجا که گروه‌های آزمایشی و کنترل، در یک دوره زمانی واحد موتورسیکلت رانده و دچار سانحه شده بودند، احتمال اینکه یک رویداد تاریخی همبسته با مداخله، سبب کاهش جراحات جدی در گروه موتورسیکلت‌های کوچک شده باشد، ناچیز است. چنین رویدادی باید باعث کاهش تصادفات گروه کنترل نیز شده باشد. وجود امکان

بررسی وجود تهدید گذشت زمان یکی از مزایای مهم طرح‌سری زمانی واجد گروه کنترل است. با این وجود، اگر یکی از گروهها، مجموعه رویدادهایی را تجربه کند که گروه دیگر تجربه نکرده، آنگاه تهدید گذشت زمان محلی مشکل‌ساز می‌شود. حتی در اینصورت هم، گذشت زمان محلی تنها زمانی روایی درونی را تهدید می‌کند که رویدادهای مداخله‌گر موردنظر، همزمان با مداخله رخ داده، و اثری هم‌جهت با اثر مداخله داشته باشند. احتمال وجود چنین رویدادهایی در مورد گروههایی که بالارت و ریبا در این آزمایش مورد استفاده قرار دادند (که به میزان زیادی قابل‌قیاس با یکدیگر بودند) وجود ندارد. اما در گروه‌هایی که شباهت کمتری با یکدیگر دارند، احتمال تأثیر گذشت زمان محلی افزایش می‌یابد.

به کمک سریهای کنترل دستکاری‌نشده می‌توان دیگر تهدیدهای روایی درونی که تنها یک سری‌زمانی منفرد را تحت تأثیر قرار می‌دهند، را بررسی کرد. به عنوان مثال، در مورد بالارت و ریبا، ابزار اندازه‌گیری بین گروه‌های کنترل و درمان، و همچنین قبل و بعد از اجرایی شدن قانون، مشابه است. به نظر می‌رسد که تمامی گروه‌ها، قبل از مداخله با یک آهنگ تغییر می‌کنند (یعنی بلوغ مشابه). مداخله درست بعد از مشاهده یک مقدار حدی (ماکزیمم یا مینیمم) خاص انجام نشده است، بنابراین احتمال وجود تهدید رگرسیون بعید به نظر می‌رسد. با این وجود، از آنجا که سری‌زمانی گروه کنترل به صورت غیرتصادفی تشکیل شده است، سوگیری انتخاب می‌تواند یکی از مشکلات بالقوه باشد، اگرچه معمولاً این تهدید چندان موجّه نیست. به عنوان مثال، در مورد بالارت و ریبا، شاید آن دسته از رانندگانی که بیشترین میزان نگرانی را در مورد ایمنی خود دارند، تا پیش از تصویب قانون، موتورسیکلت بزرگ سوار می‌شدند، چون قانون پیش از این هم رانندگان موتورسیکلتهای بزرگ را موظف به پوسیدن کلاه ایمنی می‌کرد. بنابراین، زمانی که قانون برای موتورهای کوچک تغییر کرد، آنها نیز احساس کردند که می‌توانند براحتی از این موتورها استفاده کنند. اما این احتمال نیز وجود دارد که این افراد با سطح نگرانی بالا نسبت به ایمنی، قبل از اجرایی شدن قانون هم از کلاه ایمنی استفاده می‌کرده‌اند. می‌توانیم این نکته را نیز درنظر بگیریم که شاید این رانندگان به علت فشار اجتماعی از سوی دیگر رانندگان موتورهای کوچک، ناچار بودند کلاه ایمنی خود را بردارند. اگرچه نمی‌توانیم تعیین کنیم که این تهدید انتخاب پیچیده چه اثری در شرایط مطالعه بارسلونا داشته است.

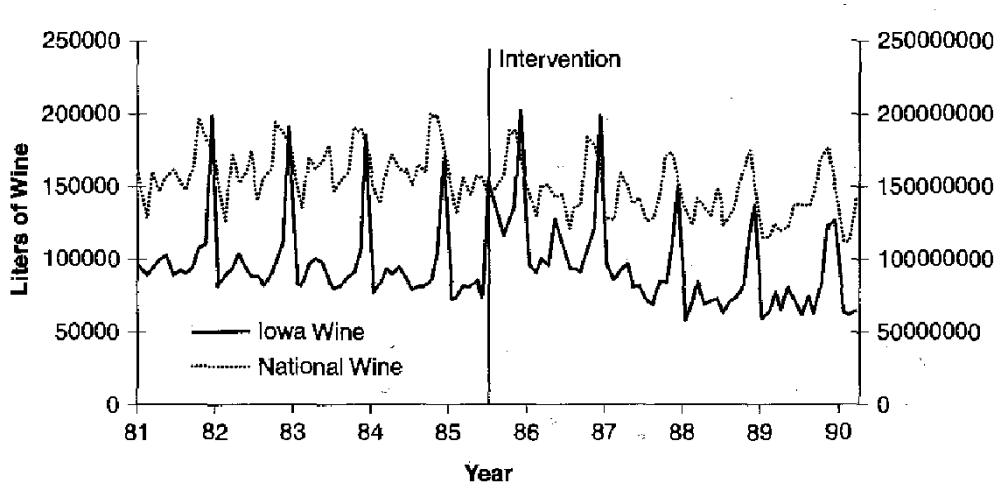


شکل ۶ – ۴. اثر اجبار کلاه ایمنی بر روی جراحت‌های جدی ناشی از موتورسیکلت

"Impact of legislation requiring moped and motorbike riders to wear helmets," by X. Ballart and C. Riba. 1995, Evaluation and Program Planning, 18, pp. 311-320. Copyright 1995 by Elsevier Science Ltd.

یک سری‌زمانی قطع‌شده دیگر (با ۱۱۱ مشاهده) با یک گروه کنترل (شکل ۶.۵)، آثار تغییر در قوانین مصرف نوشیدنی در ایالت آیوا مصوب سال ۱۹۸۵ را، روی فروش کلی نوشیدنی در این ایالت مورد بررسی قرار داد (Mulford, Ledolter & Fitzgerald, 1992). در اول جولای ۱۹۸۵، به انحصار دولت برای فروش نوشیدنی پایان داده شد. قبل از آن، تنها حدود ۲۰۰ مغازه ایالتی اجازه فروش نوشیدنی را داشتند. بعد از آن، مغازه‌های فروش نوشیدنی بخش خصوصی هم مجوز گرفتند، و به سرعت حدود ۱۲۰۰ مغازه خصوصی باز شد. برخی از افراد نگران شدند که افزایش دسترسی، منجر به افزایش مصرف الکل و به تبع افزایش اثرات منفی آن گردد. یک تحلیل سری‌زمانی زودهنگام فروش الکل طی مدت ۲.۵ سال بعد از اجرایی شدن قانون (تا دسامبر ۱۹۸۷) را موردبررسی قرار داد. یافته‌ها نشان از افزایش مصرف نوشیدنی به میزان ۹۳٪ داشت (Wageuaar & Holder, 1991). ملفورد و همکارانش (Mulford et al., 1992) با افزودن یک سری‌زمانی کنترل، و توسعه داده‌ها از نظر زمانی تا سال ۱۹۹۰، این مسأله را مورد بررسی بیشتر قرار دادند. شکل ۶.۵ داده‌های آنها را از ۱۹۸۱ تا اوایل ۱۹۹۰ که تقریباً ۵ سال بعد از اجرایی شدن قانون است، نشان می‌دهد. آنها از داده‌های فروش نوشیدنی در سطح ملی طی همان سالها، به عنوان کنترل استفاده کردند. سری‌زمانی فروش نوشیدنی ایوا بیانگر افزایش مصرف نوشیدنی با اجرایی شدن قانون بود. در واقع، مشخص شد که تا ۱۹۸۸ یعنی شش ماه بعد از اینکه واگنار و هلدر جمع‌آوری داده را متوقف کرده بودند، این الگوی افزایش در فروش ادامه داشته، اما بعد از آن، فروش نوشیدنی در ایوا به سطح پیش از مداخله برگشته است. ملفورد و همکارانش (Mulford et al., 1992) توانستند نشان‌دهند که افزایش فروش موقت بوده، و تا حدودی به

این علّت است که ۱۲۰۰ مغازه جدید می‌خواستند قفسه‌های خود را از محصول پر کنند، نه اینکه مصرف مشتریان افزایش پیدا کرده باشد.



شکل ۶.۵. اثر قانون آزاد شدن مغازه‌های بخش خصوصی در ایوا روی فروش نوشیدنی. از داده‌های ملّی به عنوان کنترل استفاده شده است.

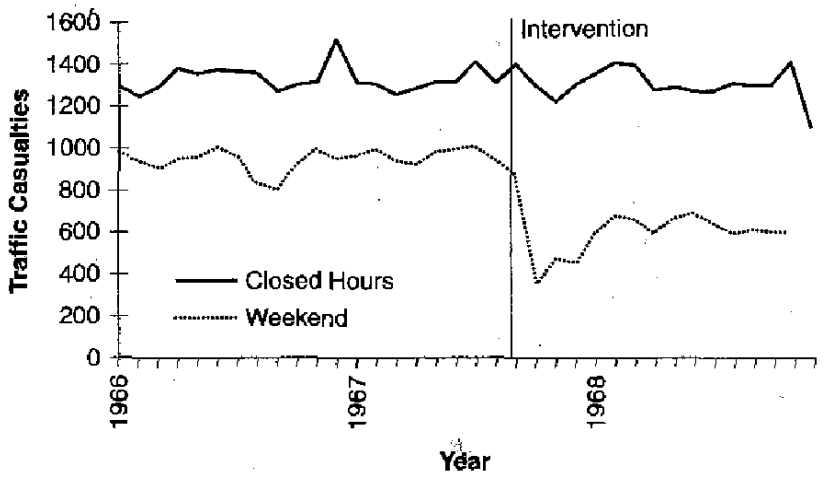
"Alcohol Availability and Consumption: Iowa Sales Data Revisited" by H. A. Mulford, J. Ledolter and J. L. Fitzgerald, 1992, Journal of Studies on Alcohol, 53, pp. 487-494. Copyright 1992 by Alcohol Research Documentation, Inc., Rutgers Center of Alcohol Studies, Piscataway NJ08855.

زمانی که این مغازه‌های جدید کاملاً قفسه‌های خود را پر کردند، فروش به سطح نرمال بازگشت. نکته‌ای که از این مثال می‌توان دریافت این است که شاید برای تعیین پایداری زمانی یک اثر، نیاز به یک سری‌زمانی طولانی‌مدت باشد. افزودن کنترل ملّی در مطالعه مولفورد و همکارانش (۱۹۹۲) سبب حذف اثر گذشت زمان شده، و ارزیابی پایداری زمانی اثرات مداخله را آسانتر کرده است.

افزودن متغیرهای وابسته غیرهم‌ارز

با گردآوری داده‌های سری‌زمانی برای یک متغیر وابسته، که مداخله باید روی آن تأثیر بگذارد، و یک متغیر وابسته غیرهم‌ارز که مداخله نباید روی آن تاثیر بگذارد، اما پاسخ آن به تهدید روایی دائم، شبیه متغیر وابسته اولیه است، می‌توان تهدیدهای مختلف روایی درونی در سری‌زمانی را بررسی کرد، و اعتبار ساختاری اثر را افزایش داد. دو متغیر وابسته باید از نظر مفهومی مرتبط باشند. نمودار طرح به صورت زیر است:

O_{A1}	O_{A2}	O_{A3}	O_{A4}	O_{A5}	X	O_{A6}	O_{A7}	O_{A8}	O_{A9}	O_{A10}
O_{B1}	O_{B2}	O_{B3}	O_{B4}	O_{B5}	X	O_{B6}	O_{B7}	O_{B8}	O_{B9}	O_{B10}



شکل۶.۶. اثر سخت‌گیری در الکل‌سنجی در بریتانیا روی مشکلات ترافیکی در شب‌های آخر هفته که رستورانها باز هستند، در مقایسه با زمانی که رستورانهاها بسته هستند.

"Determining the social effects of a legal reform: The British 'breathalyser' crackdown of 1967," by H. L. Ross, D. T. Campbell, and G. V. Glass, 1970, American Behavioral Scientist, 19, pp. 493-509. Copyright 1970 by Sage Publications.

در این نمودار، سری مشاهداتی A، متغیر وابسته مدنظر را نشان می‌دهد، و سری مشاهداتی B، متغیر وابسته غیرهم‌ارز را نشان می‌دهد.

مک سوئینی (McSweeny, 1978) در مثال راهنمای تلفن سینسیناتی‌بل، از یک متغیر وابسته غیرهم‌ارز بهره گرفت، اگرچه نمودار آن در شکل ۶.۱ نشان داده نشده است. هزینه اعمال‌شده بر راهنمایی تلفنی مربوط به راهنماییهای تلفن محلی بود، نه راهنمایی در مورد تلفنهای راه دور. اگر این اثر، در نتیجه اعمال هزینه بود، تنها تماس‌های محلی باید تغییر پیدا می‌کرد؛ اما اگر این اثر در نتیجه رویدادهای مرتبط با گذشت زمان دیگری بود که بر همه تماس‌ها با راهنمای تلفنی تأثیرگذار بودند، آنگاه تماس با راهنمای تلفن راه دور نیز باید همزمان با تغییرات سری‌زمانی راهنمای تلفن محلی، و به همان میزان، تغییر می‌کرد. مک سوئینی، هر دو سری‌زمانی را رسم کرد و دریافت که تنها تماس با راهنمای تلفن محلی تغییر کرده، و تماس‌های راه دور تغییری نکرده‌اند.

مثالی دیگر از این طرح، مربوط به مطالعه تأثیر تشدید الکل‌سنجی در بریتانیاست (Ross, Campbell & Glass, 1970؛ شکل ۶.۶). این سری‌زمانی، تنها ۳۵ مشاهده داشت، اما اثر آن بسیار قابل‌توجه بود. الکل‌سنجی برای

شناسایی رانندگان مست به کار می‌رود، و در نتیجه، خطر تصادفات شدید را کاهش می‌دهد. تحت قوانین مصرف الکل بریتانیا در آن زمان، رستورانهای مجاز به فروش الکل تنها در مدت زمان محدودی در روز می‌توانستند باز باشند. اگر بخش بزرگی از مشکلات ترافیکی ناشی از نوشیدن الکل در رستوران (در مقایسه با خانه) بود، باید الکل‌سنجی، تصادفات ترافیکی شدید را در ساعات مشخصی از روز یا شب‌های پایان هفته کاهش می‌داد؛ یعنی همان زمانهایی که رستورانها پر از مشتری بودند. همچنین در زمان‌های دیگر که رستورانها بسته بودند، الکل‌سنجی نمی‌بایست تأثیری روی تصادفات جاده‌ای داشته باشد. در واقع، شکل ۶.۶ افت بسیار شدید میزان تصادفات در پایان هفته (خروجی مدنظر) را در زمان مداخله نشان می‌داد، اما برای زمانهایی که رستورانها بسته بودند، کاهش اندک یا نزدیک به صفر مشاهده شد (متغیر وابسته غیرهم‌ارز). تحلیل آماری، این کاهش را تأیید می‌کرد.

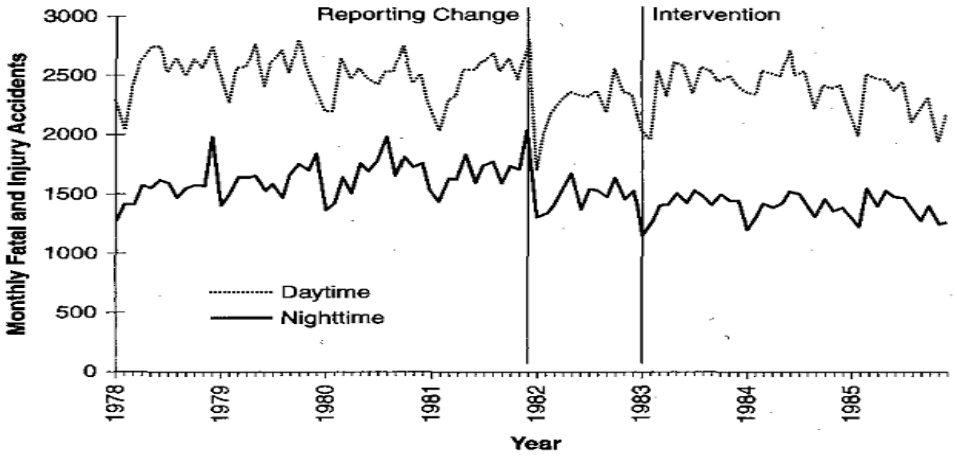
تفکیک تصادف‌هایی که در زمان بسته بودن و باز بودن رستوران‌ها صورت گرفته‌اند، حائز اهمیت است، زیرا اغلب تهدیدهای مربوط به گذشت زمان برای کاهش تصادفات جدّی، بر همه تصادفات جدّی - فارغ از زمان شبانه‌روز- تاثیر می‌گذارند. در مورد تغییرات آب‌وهوایی، تولید ماشین‌های ایمن‌تر، سخت‌گیری پلیس روی سرعت، گزارش همزمان روزنامه‌ها از نرخ بالای تصادفات، و به ویژه تصادفات ناخوشایند، و غیره، نیز این مسأله صادق است. بنابراین یافتن مشکلات مربوط به روایی درونی یا نتایج آماری در این داده‌ها، دشوارتر است.

با این وجود، می‌توان سؤالاتی را راجع به روایی بیرونی طرح کرد. به عنوان مثال، آیا در ایالات متحده نیز نتایج مشابهی بدست آمده است؟ نگرانی دیگر این است که آیا این اثرات روی یک گروه از رانندگان شدیدتر از رانندگان دیگر است؟ مشکل دیگر مربوط به آثار جانبی پیش‌بینی نشده الکل‌سنج‌هاست. این مسأله بر نرخ بیمه تصادفات، فروش نوشیدنیها، اطمینان عمومی از نقش ابتداعات فناورانه در حل مشکلات اجتماعی، فروش گجت‌های فنی به پلیس، یا نحوه تعامل دادگاه با رانندگان مست چه تاثیری دارد؟ راس (Ross, 1973) به بررسی این مسائل پرداخت. شکل ۶.۶ نشان می‌دهد که بخشی از کاهش‌های اولیه تصادفات جدی در آخر هفته، به مرور زمان از بین رفته است. در ابتدا، نرخ تصادف کاهش می‌یابد، اما دوباره به همان سطح سری‌زمانی کنترل باز می‌گردد. بنابراین این اثر، تنها یک اثر کاهشی موقت و جزئی در تصادفات شدید بود.

راس همچنین بیان کرد که الکل‌سنجی به صورت گسترده‌تر در کشور معرفی شده بود. آیا عمومیت سبب می‌شد که مردم از سودمندی عدم‌رانندگی بعد از مصرف الکل مطلع شوند؟ آیا سبب می‌شد که پلیس در کنترل سرعت به ویژه در زمانی که رستورانها باز هستند، هشیارتر عمل کند؟ آیا تعداد کل ساعات رانندگی را کاهش می‌داد؟ آیا باعث کاهش مصرف الکل می‌شد؟ آیا سبب می‌شد که رانندگان مست، با دقت بیشتری رانندگی کنند؟ راس به شکل هوشمندانه‌ای، برخی از این تبیینها را در مطالعه بی‌اثر کرد (حذف کرد). او از نتایج بررسی‌های معمول روی مسافت رانندگی که توسط آزمایشگاه تحقیقات جاده‌ای بریتانیا انجام شده بود بهره گرفت، و نشان داد که کاربرد الکل‌سنجی، حتی با کاهش تصادفات برای هر مایل رانندگی هم رابطه دارد. البته نمی‌توان گفت که اثر الکل‌سنجی ناشی از کاهش مسافت رانندگی است. راس، فروش نوشیدنی‌ها قبل و بعد از کاربرد الکل‌سنج را نیز مطالعه کرد، و شواهدی از ناپیوستگی بعد از کاربرد الکل‌سنج مشاهده نکرد.

بنابراین، این تفسیر که الکل‌سنج، مصرف الکل را کاهش می‌دهد را کنار گذاشت. او همچنین توانست نشان دهد که ۱۰ ماه بعد از به کارگیری الکل‌سنج، نسبت به ده ماه قبل از کاربرد الکل‌سنج، افراد مست بیشتری پیاده به سوی منزل می‌رفتند. در نهایت، او نشان داد که فوت‌شده‌های پس از کاربرد الکل‌سنج، نسبت به فوت‌شده‌های قبل از آن، سطح الکل کمتری در خون خود داشتند. این تحلیل‌ها بیانگر این است که کاهش تعداد افراد بسیار–مست در حال رانندگی توضیح درست است، و نه کاهش میزان مصرف یا مسافت رانندگی. داده‌هایی که راس استفاده کرد تا بخشی از توضیحات مربوط به ساختار علّی را کنار بگذارد، بیانگر اهمیت انجام این کار، دشواری و هزینه بالای انجام آن و وجود تعداد زیادی توضیحات بی‌ربط برای تأثیر یک اقدام جدید است.

در نهایت مشکل راس این بود که توضیح دهد چرا آثار الکل‌سنج دائمی نیستند. تحلیل او نشان داد که دادگاه‌های بریتانیا نتوانستند تنبیه مناسبی برای رانندگان مست شناسایی‌شده اعمال کنند، و در نتیجه قدرت این قانون از دست رفته است. بنابراین، استنباط نهایی راس بسیار سودمند بود: اگر الکل‌سنج برای محدود کردن رانندگی افراد مست به کار گرفته شود، به کاهش تصادفات جاده‌ای شدید کمک خواهد کرد، اما تنها زمانی مفید خواهد بود که دادگاه‌ها، قانونی راجع به رانندگی در حین مستی وضع کنند.



شکل ۶ – ۷. اثر قانون رانندگی در حین مستی بر اساس تغییر در گزارش تصادفات.

"The impact of drunk driving legislation in Louisiana," M. W. Neustrom and W. M. Norton, 1993, Journal of safety research, 24, pp. 107-121. Copyright by Elsevier Science.

نوستروم و نورتون (Neustrom and Norton, 1993) یک سری‌زمانی با ۹۶ مشاهده (شکل ۶.۷) در مورد اثرات قانون رانندگی در مستی مصوب سال ۱۹۸۳ در لوئیزیانا ارائه کردند. نوسترون و نوترون (۱۹۹۳) این فرضیه را ارائه کردند که اثر این قانون در شب قوی‌تر از روز خواهد بود، زیرا تحقیقات گذشته نشان داده است که تصادفات مرتبط با الکل در شب جدی‌تر و مکررتر رخ می‌دهند. با این وجود، کمی قبل از اجرای این قانون به دلیل کمبود

منابع، کمبودی موقتی در پرسنل پلیس رخ داده بود که این کمبود همزمان با دوره اجرایی شدن قانون رانندگی درمستی برطرف شده بود. ممکن است این رویداد باعث شده باشد افسرهای پلیس، گزارش‌های بیشتری در مورد تصادفاتی بنویسند که قانون جدید قرار بوده آنها را کاهش دهد. در نتیجه، این دو اثر متقابل ممکن است نتیجه همدیگر را خنثی کرده باشند. همانگونه که شکل ۶.۷ نشان می‌دهد، سری‌زمانی کنترل برای روشنتر شدن این اثرات مداخله‌کننده مفید بود. در سری‌زمانی مرتبط با تصادفات در طول روز، گزارش تصادفات بعد از تصویب قانون جدید افزایش یافته است، اما در سری‌زمانی شب این تصادفات اندکی کاهش‌یافته است. نوسترون و نورتون (۱۹۹۳) با در نظر گرفتن افزایش گزارش تصادفات که با رفع کمبود پرسنل رخ داده است، برآورد کردند که قانون جدید سبب کاهش ۳۱۲ تصادف در سری‌زمانی شب، و کاهش ۱۲۴ تصادف در سری‌زمانی روزانه شود. بنابراین، اگر زمان شروع تغییر در گزارش‌نویسی مشخص باشد، و اگر طرح، متغیر وابسته غیرهم‌ارز داشته باشد، می‌توان اثرات تغییر در گزارش‌نویسی را برآورد کرد (در این مورد، روز در مقابل شب متغیر تصادفی غیرهم‌ارز است، که فرض می‌شود اثر آن در شب قوی‌تر باشد).

حذف مداخله در زمانی معین

تأثیر مداخله را نه تنها با آثاری که در حین اجرای مداخله رخ می‌دهد، بلکه با آثاری که بعد از حذف مداخله مشاهده می‌شود، نیز می‌توان شرح داد. طرح مداخله حذف‌شده در اینجا رسم شده است که در آن، X، مداخله و ~~X~~ حذف آن را نشان می‌دهد.

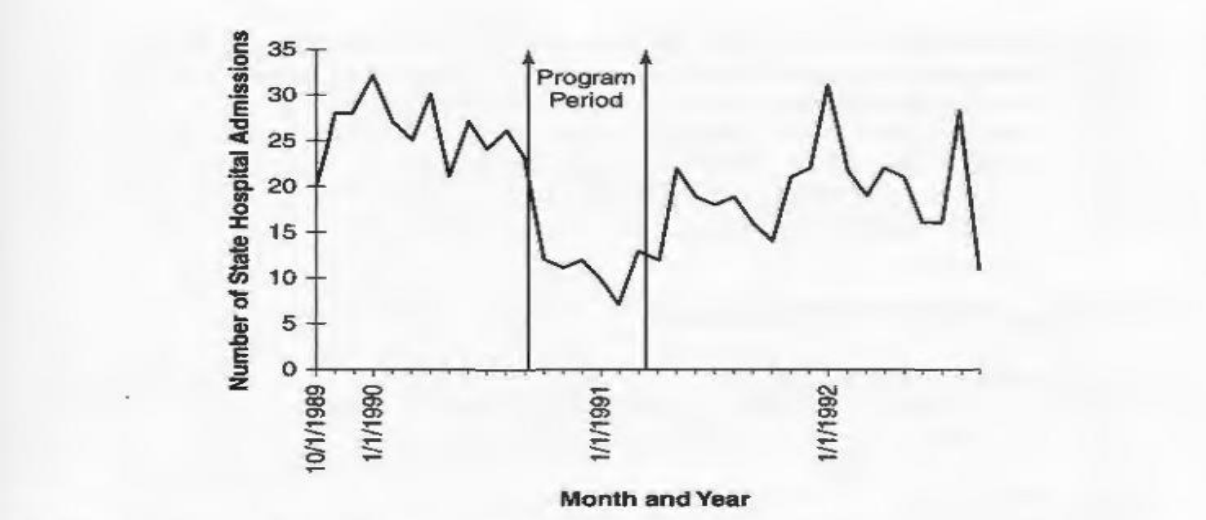
O_1	O_2	O_3	O_4	X	O_5	O_6	O_7	O_8	O_9	~~X~~	O_{10}	O_{11}	O_{12}	O_{13}

این طرح همانند داشتن دو سری‌زمانی متوالی قطع‌شده ساده است. اولی که در اینجا از O_1 تا O_9 است، اثر اعمال مداخله را بررسی می‌کند، و دومی که از O_5 تا O_{13} است، اثر حذف مداخله موجود را بررسی می‌کند. تفسیرپذیرترین الگوی اثرات این است که سطح یا شیب بین O_4 و O_5 در یک جهت تغییر می‌کنند، و سپس بین O_9 و O_{10} در جهت مخالف تغییر می‌کنند.

شکل ۶.۸ مثالی از مطالعه‌ای با ۳۶ مشاهده را نشان می‌دهد که توسط ردینگ و رافلسون (Reding and Raphelson, 1995) انجام گرفت. در اکتبر ۱۹۸۹، یک روانپزشک به تیم مداخله بحران روانپزشکی سیار اضافه شد تا این تیم بتواند مداخلات روانپزشکی لحظه‌ای فوری در محل انجام دهد، به این امید که این کار از پذیرش افراد در بیمارستان دولتی پیشگیری نماید. این اثر حاصل شد. شش ماه بعد، چندین عامل منجر شدند که این روانپزشک تیم حذف شود. اینکار باعث افزایش مجدد پذیرشها در بیمارستان دولتی شد. اگرچه در شکل ۶.۸ نشان داده نشده است، اما

نویسندگان بعدها استنباط علّی خود را از طریق مقایسه این نتایج با داده‌های پذیرش یک بیمارستان روانپزشکی خصوصی محلی در همان زمان که تغییری در پذیرش آن مشاهده نشده بود، تقویت کردند. در اینجا، سه عنصر طراحی استنباط علّی را تسهیل می‌کردند: سری‌زمانی قطع‌شده، حذف مداخله، و بیمارستان‌های خصوصی به عنوان کنترل.

حذف مداخله مزایای زیادی برای طرح سری‌زمانی دارد. یکی از این مزایا این است که تهدید گذشت زمان کاهش می‌یابد، زیرا تنها آن دسته از تهدیدهای گذشت زمان اهمیت دارند که یا در جهات مختلف و در زمان‌های مختلف عمل می‌کنند، و یا آنهایی که دربرگیرنده دو نیروی زمانی مختلف هستند که در دو جهت متفاوت در زمان‌های متفاوت عمل می‌کنند، و این زمانها به طور اتفاقی مقارن با زمان اعمال و حذف مداخله بوده است. انتخاب و ریزش تهدیدی به حساب نمی‌آیند، مگر اینکه افراد متفاوتی در نقاط زمانی مختلف در طرح وارد یا خارج شوند. خطای ابزار نیز تهدید نیستند، اگرچه اثر سقف یا کف (یعنی ممکن است شرکت‌کنندگان به کمترین یا بیشترین امتیاز ممکن برسند، بگونه‌ای که دیگر امکان تغییر بیشتر در آن جهت برای آنها وجود نداشته باشد) می‌تواند مشکلاتی ایجاد می‌کند، اگر نقطه کف یا سقف در همان نقطه‌ای ایجاد شود که مداخله در آن حذف شده است. دیگر اثرات مربوط به ابزار نیز چندان موجّه نیستند، زیرا به این معناست که ابزارهایی یکسان باید توانسته باشند در زمان‌های متفاوتی، هم باعث کاهش شده باشند و هم باعث افزایش.



شکل ۶ – ۸. اثر مداخله بحران روان‌پزشکی بر بستری‌شدن در بیمارستان

Around-the-clock mobile psychiatric crisis intervention: Another effective alternative to psychiatric hospitalization. G. R. Reding and M. Raphelson, 1995. Community Mental Health Journal, 31, pp. 179-187. Copyright 1995 by Kluwer Academic Publishers.

سری کنترل بیمارستان خصوصی به حذف اثر گذشت زمان (یعنی احتمال اینکه یک رویداد دیگر، در حالت کلی در پذیرش بیمارستان تأثیرگذار باشد)، و روندهای بلوغی مانند الگوهای چرخه‌ای پذیرش در بیمارستان که با تغییر فصل بوجود می‌آیند، نیز کمک کرده است. تهدید اول را می‌توان به آسانی از طریق مصاحبه کیفی با افراد آگاه بیمارستان مثل رئیس بیمارستان

بررسی کرد. تهدید دوم محتمل‌تر است، زیرا این پدیده‌ای شناخته‌شده است که برخی بیماران به گونه‌ای برنامه‌ریزی می‌کنند که در فصل سرد زمستان از بیمارستان مرخص شوند تنها برای اینکه در طول تابستان به این مناطق خنک‌تر بازگردند.

بر اساس این طرح، و اثرات مشاهده‌شده در دو جهت مخالف در X و ~~X~~، گاهی می‌توان گفت که ناپدیدشدن اثر اولیه به دلیل حذف مداخله نیست، بلکه به علت تضعیف روحیه با رنجش و ناراحتی ناشی از حذف مداخله است. در مقاله ردینگ و رافلسون (Reding and Raphelson, 1995) مشاهده می‌شود که بعد از حذف روانپزشک از گروه، سطح پذیرش بیمارستان بالاتر رفته است، که این امر می‌تواند ناشی از تضعیف روحیه باقیمانده اعضای تیم به علت حذف یک مداخله موفق باشد. اگر اینگونه باشد، طرحهای حذف مداخله احتمالاً زمانی تفسیرپذیرتر هستند، که مطلوبیت مداخله برای شرکت‌کنندگان کمتر باشد. با این وجود، این تضعیف روحیه، منجر به تهدید اثر اعمال مداخله نشده است. زمانی که طرح مداخله حذف‌شده نتایجی در جهات مختلف تولید می‌کند، دو تفسیر جایگزین *متفاوت* نیز لازم خواهد بود تا بتوان استنباط علّی را باطل کرد. در نهایت، طرح‌های مداخله حذف‌شده تنها زمانی مفید هستند، که حذف مداخلهها اخلاقی باشد، و همینطور اثرات مداخله گذرا باشد، و با حذف مداخله این اثرات از بین بروند. بنابراین این طرح را نمی‌توان برای آن دسته از مداخلهها که اثرات بلندمدتی دارند به کار گرفت.

افزودن چندین تکرار

می‌توان طرح پیشین را به این صورت بسط داد که یک مداخله وارد شود، حذف شود، دوباره وارد شود، دوباره حذف شود، و این کار مطابق با جدول زمانبندی شده ادامه یابد. نمودار طرح به صورت زیر است:

O_1	O_2	X	O_3	O_4	~~X~~	O_5	O_6	X	O_7	O_8	~~X~~	O_9	O_{10}	X	O_1	O_{12}	~~X~~	O_{13}	O_{14}

اگر متغیر وابسته با هر بار اعمال و حذف مداخله، به‌طور مشابهی رفتار کند، اما جهت پاسخ‌ها برای ورود و حذف متفاوت باشد، مداخله موثر فرض می‌شود. این طرح برای ارزیابی اثر مداخلات روانشناسی (Marascuilo & Busk, 1988؛ Wampold & Worsham, 1986) یا پزشکی (Weiss et al., 1980) روی بیماران مختلف مورد استفاده قرار گرفته است. به عنوان مثال، مک‌لود، تیلور، کوهن و کولن (Mcleod, Taylor, Cohen and Cullen, 1986) اثر یک دارو را با اثر دارونما، بر درمان بیماران مبتلا به التهاب ایلئوستومی قابل‌کنترل مقایسه کردند. دارو و دارونما به صورت تصادفی تخصیص داده شدند (و طرح دوبرابر-کور[۳۶۰] بود، اگرچه کوری در این طرح بسیار دشوار است)، و این کار در ده دوره مطالعاتی که هر کدام ۱۴ روز بودند به طول انجامید (اگرچه داده‌ها به صورت روزانه گردآوری

[360] Double blind

شدند، اما تنها به صورت تجمیع‌شده برای ۱۰ دوره مداخله در دسترس بودند). بیماران نتایج درمان خود را در غالب متغیرهایی مانند تهوع، شکم درد، باد شکم، مدفوع حجیم، مدفوع شل، و بوی بد، گزارش می‌کردند. شکل ۶.۹، نتایج مربوط به حال خوب و درد را نشان می‌دهد، که هر دو بیانگر تأثیر درمان هستند. زیرا در اثر درمان، حالت تهوع کاهش، و حال خوب افزایش یافته است. اگرچه اثر بر درد بسیار قویتر از اثر بر حال خوب بود.

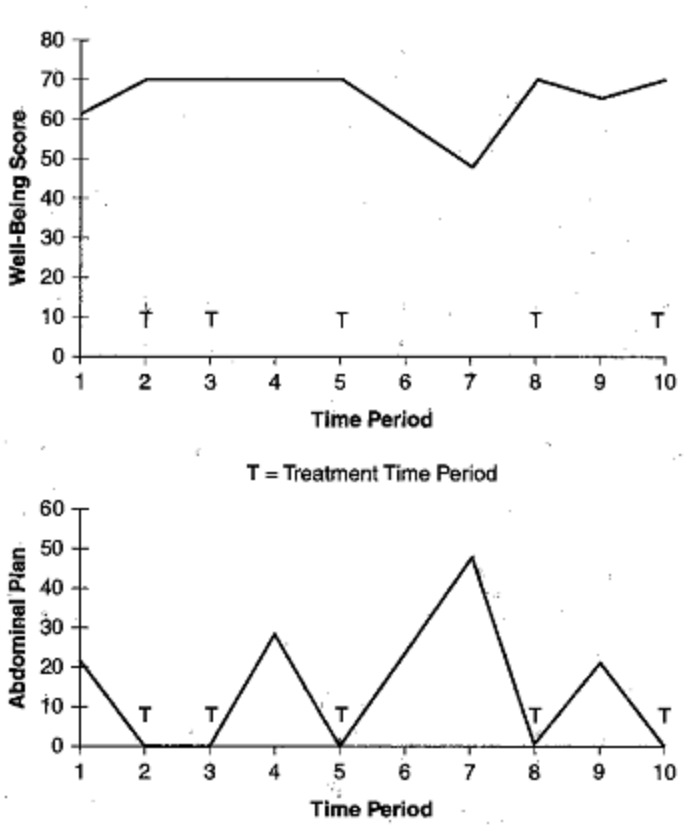
یکی از مشکلات این طرح، زمانبندی اجرای مداخله و حذف آن است. اگرچه زمانبندی معمولاً به صورت سیستماتیک، و یا در پاسخ به وضعیت علائم بیمار انجام می‌شود (Barlow & Hersen, 1984)، بسیاری از مزایای این طرح برآمده از زمانبندی تصادفی درمانهاست، البته به صورتی که تناوب اعمال X و ~~X~~ حفظ شود (Edgington, 1987, 1992). زمانبندی تصادفی، تهدید بلوغ چرخه‌ای را بلاموضوع می‌کند (در این بلوغ فرض می‌شود که سری‌زمانی، یک چرخه منظم بالا و پایین را حتی در غیاب مداخله تجربه می‌کند). با انجام دیگر اصلاحات می‌توان محدوده کاربرد این طرح را افزایش داد. به عنوان مثال، می‌توان این طرح را برای مقایسه دو درمان مختلف که از لحاظ نظری مناسب بوده‌اند به کار برد؛ به این صورت که X_1 جایگزین X شده و X_2 جایگزین ~~X~~ می‌شود. همچنین می‌توان از یک عامل مداخله توأم[۳۶۱] $(X_1 + X_2)$ بهره برد. اولیری، بکر، ایوانز و سادارگاس(O'leary, Becker, Evans and Saudargas, 1969) اینکار را انجام دادند. آنها رفتار تخریبی هفت کودک را در یک کلاس که تحت‌تأثیر مداخلههایی مشتمل بر اول دادن قانون، سپس افزودن ساختار آموزشی به قوانین، پس از آن استفاده از قانون، ساختار آموزشی و تشویق در ازای انجام رفتار خوب، یا چشم‌پوشی از رفتار بد دیگران، و پس از آن اضافه کردن یک نظام ژتونی[۳۶۲] به همه اینها بود، مورد بررسی قرار دادند. آنگاه برای نشان دادن کنترل روی کل پدیده‌ها، همه مداخلهها حذف شده، و دوباره اجرا شدند. در یک حالت متفاوت دیگر از این طرح، مداخله با شدت‌های مختلف (معمولاً افزایشی) انجام می‌شود تا رابطه پاسخ و دوز بررسی شود؛ برای مثال، هارتمان و هال (Hartmann and Hall, 1976) اثر مجازات‌های بطور فزاینده بالا را بر افزایش در تعداد سیگارهای مصرف‌شده در هر روز بررسی کردند.

این طرح در عمل محدودیت‌های قابل‌توجهی دارد. اول اینکه، همانند طرح مداخله حذف‌شده، تنها زمانی می‌توان این طرح را اجرا کرد که انتظار برود اثر مداخله به سرعت کاهش یافته، و ناپدید شود. همچنین، این طرح نیازمند میزان معینی کنترل آزمایشی است که اجرای آن در خارج از مختصات آزمایشگاهی به ندرت امکان‌پذیر است، از سوی دیگر، به زمینه مداخله با طراحی تک‌موردی خاص، یا شرایط سازمانی بسته، مانند مدرسه یا زندان نیاز دارد. با این وجود، بارلو و هارسن (Barlow and Hersen, 1984) بحث کاملی درمورد گزینه‌های طراحی مختلف در شرایطی که امکان اجرای این طرحها وجود داشته باشد، ارائه کرده‌اند.

---

[361] Joint treatment factor

[۳۶۲] نظامی که در آن با کنترل دقیق تقویت ها ، رفتارهای مطلوب رواج داده می شوند. ایجاد چنین نظامی مستلزم مشخص کردن تقویت کننده ای فوری برای هر رفتار و پشتوانه های آن تقویت کننده هاست که بیماران می توانند تقویت کننده ها را با آنها مبادله کنند.



شکل ۶.۹. اثر درمان روی التهاب ایلئوستومی قاره‌ای. در گراف‌ها، حرف T، بیانگر دوره زمانی‌ای است که درمان در آن صورت گرفته است.

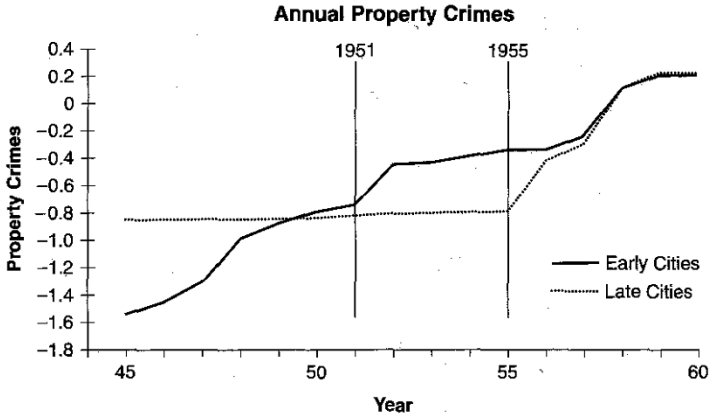
"Single patient randomized clinical trial: Its use in determining optimal treatment for patient with information of a Kock continent ileostomy reservoir," by R. S. McLead et al., 1986, Lancet,1, pp. 726-728. Copyright 1986 by The Lancet Publishing Group.

افزودن تکرارهای جابجاشونده

دو گروه غیرهم‌ارز (یا بیشتر) را تصور کنید که هر کدام، مداخلاتی را در زمان‌های مختلف و با ترتیبی متغیر دریافت می‌کنند، به گونه‌ای که (۱) زمانیکه یک گروه مداخله را دریافت می‌کند، گروه دیگر به عنوان کنترل عمل می‌کند، و (۲) زمانی که گروه کنترل در ادامه فرایند، مداخله دریافت می‌کند، گروه مداخله اصلی، به عنوان کنترل دستکاری شده عمل می‌کند. این طرح را می‌توان به صورت زیر رسم کرد:

O_1	O_2	O_3	X	O_4	O_5	O_6	O_7	O_8		O_9	O_{10}	O_{11}
O_1	O_2	O_3		O_4	O_5	O_6	O_7	O_8	X	O_9	O_{10}	O_{11}

این طرح، اغلب تهدیدهای روایی درونی را کنترل می‌کند، و روایی سازه و بیرونی را نیز افزایش می‌دهد. روایی بیرونی به این دلیل افزایش می‌یابد که می‌توان یک اثر را با دو جمعیت و در دو لحظه متفاوت از زمان و گاهی در دو مختصات مختلف نشان داد. عناصر غیرمرتبط بسیاری ممکن است در جریان اجرای هر مداخله وجود داشته باشند، و اگر اندازه‌گیریها نامحسوس باشند، نگرانی خاصی درمورد برهم‌کنش مداخله با آزمون وجود نخواهد داشت.



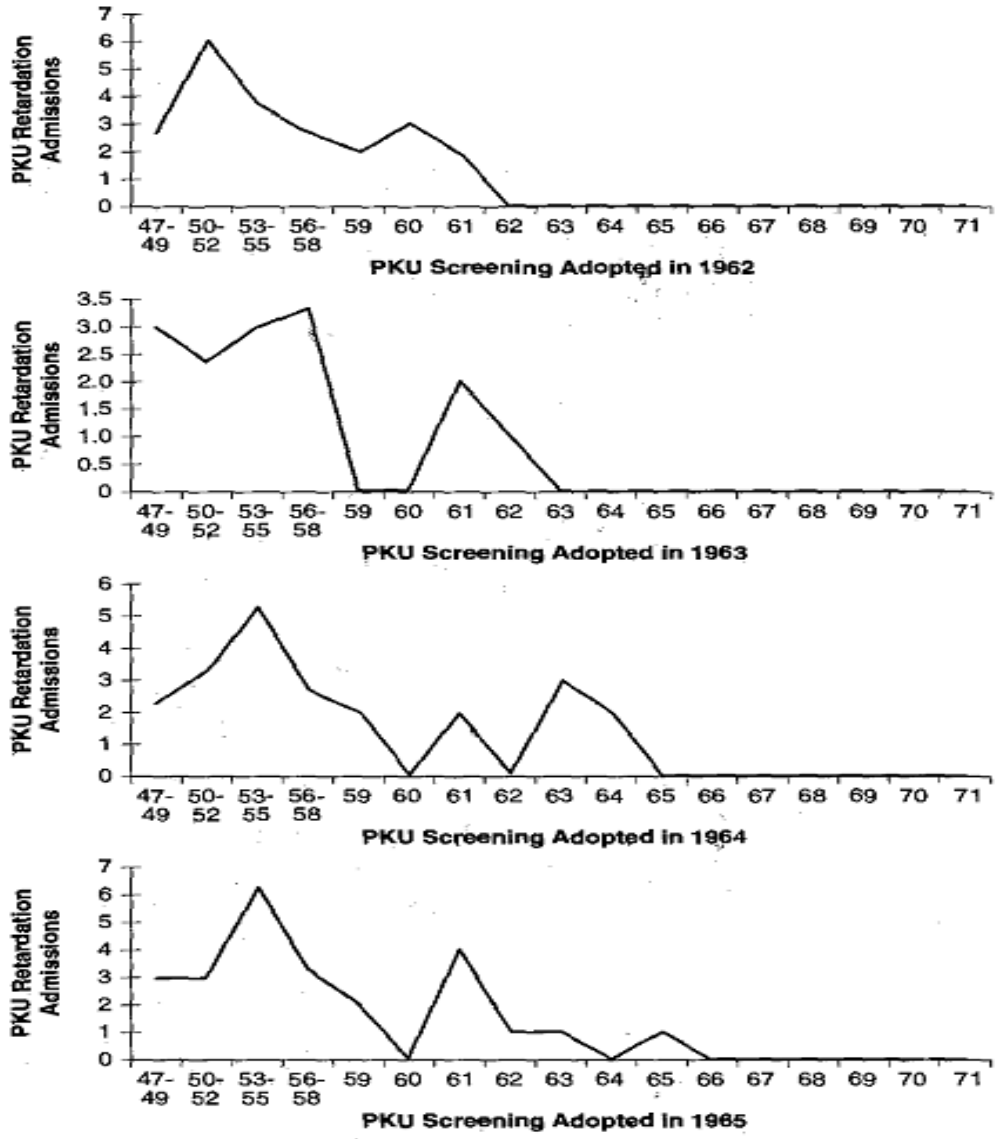
شکل ۶.۱۰. اثر ورود تلویزیون بر نرخ جرایم مربوط به اموال در شهرهایی که تلویزیون در سال‌های ۱۹۵۱ تا ۱۹۵۵ در آنها وارد شده است.

The evolution of the time series experiment by R. D. McCleary, 2000, Research design: Donals [a]
Campbell's, vol 2, edited by L. Bickman, Thousand Oaks, Ca: Sage. حق تکثیر ۲۰۰۰ توسط انتشارات Sage.

شکل ۶.۱۰ نتایج مطالعه‌ای را ارائه می‌دهد که در آن، نرخ جنایت سالانه (باتأخیر و استانداردشده) برای ۳۴ شهر که تلویزیون در سال ۱۹۵۱ به آنجا وارد شده بود، و ۳۴ شهر دیگر که تلویزیون در سال ۱۹۵۵ به آنها وارد شد ترسیم شده است (McCleary, 2000 ؛Henningan et al., 1982)، و شکاف موجود در نمودار به علّت ممنوعیت صدور مجوز تلویزیون توسط FAA در فاصله سال‌های ۱۹۵۱ تا ۱۹۵۵ است. در هر دو سری‌زمانی، ورود تلویزیون منجر به افزایش جرم در سال بعد از معرفی شده است. تکرار این اثر در دو نقطه زمانی که ۵ سال از یکدیگر فاصله دارند، سبب می‌شود که احتمال وجود تهدید گذشت زمان کم باشد. به عنوان مثال، جنگ کره در سال ۱۹۵۱ شروع شد، و رفتن مردان به ارتش سبب شد که اغلب خانه‌ها بدون نگهبان باشد، و از آنها دزدی شد؛ اما این امر نمی‌تواند توضیح دهد که چرا این اثر تنها در شهرهای جلوتر رخ داد؟ زیرا جنگ می بایست تأثیر مشابهی روی شهرهای بعدی نیز می‌گذاشت. همچنین، رکود اقتصادی سال ۱۹۵۵ می‌توانسته سبب افزایش جرم و جنایت شده باشد، اما باز هم باید بر هر دو سری تأثیر می‌گذاشت. مصنوعات رگرسیونی احتمالاً می‌توانند افزایش در شهرهای اولیه را توجیه کنند، اما به نظر نمی‌رسد که بتوانند وضعیت شهرهای بعدی را نیز تبیین کنند. در حالت کلی، به ندرت

احتمال دارد یک تهدید منفرد بتواند هر دو افزایش را توضیح دهد. تهدیدهایی مانند گذشت زمان–انتخاب یا انتخاب–ابزار می‌توانند تبیینهای جایگزینی ارائه کنند؛ مثل اینکه اگر کسی می‌توانست دو رویداد تاریخی مجزا را بیابد، یکی رویدادی که در بیشتر شهرهای اول رخ داده، اما در شهرهای بعدی رخ نداده است، و دیگری رویدادی که در اغلب شهرهای ثانویه رخ داده، اما در شهرهای اولیه رخ نداده است، و هر دو این رویدادها می‌توانسته‌اند منجر به افزایش جرم شوند. اما این احتمال هم ضعیف به نظر می‌رسد.

شکل ۶.۱۱ نتایج مطالعه‌ی دیگری را که روی نوزادان غربالگری‌شده از نظر فنیل‌کتونوری (PKU)، و به منظور جلوگیری از عقب‌ماندگی ذهنی، صورت گرفت را نشان می‌دهد (MacCready, 1974). این سری‌زمانی، ۱۷ مشاهده دارد که ۴ مورد اول آن، به صورت تجمیعی از دوره‌های ۳ ساله جداگانه هستند. غربالگری PKU در سال‌های مختلف (۱۹۶۲، ۱۹۶۳، ۱۹۶۴ و ۱۹۶۵) در ایالت‌های مختلف آمریکا و کانادا، به عنوان یک عمل استاندارد در زمان تولد شناخته شده است. این موضوع امکان داشتن مؤلفه تکرار در طراحی را تسهیل می‌کند. متغیر وابسته عبارتست از تعداد پذیرش سالانه بیماران با عقب‌ماندگی ناشی از PKU در موسسات کلینیکی ایالت یا خصوصی. شکل ۶.۱۱ نشان می‌دهد که مقدار پذیرش در سال‌های بعد از اجرایی شدن غربالگری، در هر چهار سری‌زمانی، به صفر رسیده، و در حد صفر باقی مانده است. این نتایج فرضیات مطالعه را با بکارگیری اندازه‌گیریهای آرشیوی، جمعیتهای بسیار پراکنده، مقاطع تاریخی متفاوت اجرای مداخله، عناصر مختلف مرتبط با نحوه انجام مداخله، و اندازه‌گیریهای مکرر برای تعیین اینکه آیا اثر اولیه را می‌توان به تمام طول دوره زمانی تعمیم داد، تأیید می‌کردند. اما حتی طرحهای سری‌زمانی با تکرارهای جابجا شوند نیز می‌توانند از نظر روایی درونی اشکالاتی داشته باشند. هر دو نتیجه شکل ۶.۱۱ و شواهد روایت شده، بیانگر این هستند که برخی از غربالگری‌ها قبل از الزامی‌شدن این استاندارد صورت‌گرفته‌اند. مختصصان پزشکی از طریق مقالات علمی، تماس با پزشکانی که به تازگی تعلیم دیده‌اند، سخنرانی‌های کنفرانس‌ها، و بحث با همکاران به مزایای این روش غربالگری پی برده بودند. این روند عمومی در کاهش PKU حتی قبل از اینکه غربالگری به صورت رسمی اجرا شود نیز دیده شده بود (اگرچه قبل از اجرایی شدن، مقدار بیماری هرگز به صفر نرسیده بود). شاید این عقب‌افتادگی با مراقبت بهتر مادران از خود، تغذیه بهتر و خدمات پزشکی بهتر در زمان تولد، و بعد از آن کاهش یافته است. مک‌کریدی (MacCready, 1974) برای بررسی این احتمال، اطلاعاتی راجع به عقب‌افتادگی و دلایل آن در طول دهه ۱۹۵۰ تا سال ۱۹۷۲ جمع‌آوری کرد، و یافته‌های او هیچ روند سیستماتیک رو به بالا یا رو به پایینی که بتواند حتی بخشی از تغییر در مبتلایان PKU در شکل ۶.۱۱ را تبیین کند، را نشان نمی‌داد. اما استفاده از متغیرهای وابسته غیرهم‌ارز با این شیوه سبب می‌شود که این توضیحات جایگزین ناموجّه شوند.



شکل ۶.۱۱. اثر غربالگری فنیل‌کتونوری (PKU) روی مجوز عقب‌ماندگی ناشی از PKU، با اجرای غربالگری که در طول ۴ سال در مکان‌های مختلف انجام شده است.

از Admission of phenylketonuric patients to residential institutions before and after screening programs of the newborn infant. By R.A. MacCready, 1974, Journal of Pediatrics, 85, pp. 383-385. حق تکثیر ۱۹۷۴ توسط انجمن ملی پزشکی.

طرح سری‌زمانی با تکرارهای جابجاشوند می‌تواند به شناسایی اثراتی که دوره تأخیر پیش‌بینی نشده دارند، کمک کند. با این فرض که در همه گروه‌ها، تأخیر یکسانی در اثر موردنظر وجود دارد، انتظار داریم که در یک سری زودتر از دیگری، ناپیوستگی مشاهده شود. همچنین انتظار می‌رود که دوره زمانی مابین ناپیوستگیها با دوره معین

بین اجرای مداخله در گروه‌های مختلف، برابر باشد. با این وجود، همواره این امر ممکن نیست، به عنوان مثال، ممکن است سرعت اثر مداخله با توسعه فناوری‌های جدید اجرای مداخلهها، و اثربخش‌ترشدن آنها، افزایش پیدا کند. بنابراین، بهتر است که به دنبال اختلاف نسبی در زمانهایی که در آنها یک اثر مداخله در هر یک از گروهها بروز می‌کند، باشیم. طرح سری‌زمانی با تکرارهای جابجاشونده برای بررسی آثار علّی تأخیردار در مواردی که نظریه قوی برای تأخیر موردانتظار وجود ندارد، سودمند است. با این وجود، مواردی را بهتر می‌توان تفسیر کرد که اختلاف زمانی بین گروه‌های دریافت‌کننده مداخله، با دوره زمانی بین آثار ظاهر شده در هر گروه انطباق دارد (همانگونه که در شکل ۶.۱۱ دیده می‌شود).

هر گاه که طرح سری‌زمانی را بتوان با گروه کنترل بدون‌مداخله همراه کرد، طرح سری زمانی باتکرار سودمند خواهد بود. مداخلههای موفق معمولاً برای گروه‌ها یا سازمان‌هایی که نقش کنترل بدون مداخله را داشته‌اند نیز مفید خواهند بود. نمایندگان این گروه‌ها می‌توانند بررسی کنند که آیا می‌خواهند این مداخله را دریافت کنند یا خیر. به عنوان مثال، فرض کنید هانکین و همکارانش (Hankin et al., 1993؛ شکل ۶.۳) می‌توانستند به کارگیری برچسب هشدار الکل روی بطری را در کشورهای مختلف، و در طول سالهای مختلف موردمطالعه قرار دهند. با توجه به کوچکی و باتأخیر بودن اثرات، استنباط علّی می‌توانست بسیار روشن‌تر باشد.

برخی از مشکلات پرتکرار در طراحی سری‌زمانی‌قطع‌شده

همانگونه که از مثال‌های پیشین بر می‌آید، برخی مشکلات به صورت مکرر در انجام تحقیقات سری‌زمانی قطع‌شده رخ می‌دهند، که از آن جمله می‌توان به موارد زیر اشاره داشت:

- بسیاری از مداخلهها به کندی اجرا شده، و در میان جمعیت پخش می‌شوند، بنابراین، به جای اینکه مداخله به صورت یک فرایند یکباره مدلسازی شود، بهتر است به صورت یک فرایند انتشار تدریجی مدلسازی شود.
- بسیاری از آثار با تأخیر زمانی غیرقابل پیش‌بینی رخ می‌دهند، و این تأخیر در بین جمعیت‌های مختلف یا در طول زمان، متفاوت است.
- بسیاری از سری‌های داده، بسیار کوتاه‌تر از ۱۰۰ مشاهده -که حد نصاب توصیه شده برای تحلیل آماری است- هستند.
- اختصاص مکان به داده‌ها برای بسیاری از بایگان‌ها دشوار است، یا تمایلی به در اختیار گذاشتن (دادن) داده ندارند.
- داده‌های بایگانی‌شده ممکن است حاوی داده‌هایی باشند که در آنها بازه‌های زمانی بین هر نقطه از داده‌ها بلندتر از میزان موردنیاز باشد؛ برخی از داده‌ها از دست‌رفته باشند، یا عجیب به نظر برسند؛ و یا برخی از تغییرات در تعاریف، مستند نشده باشد.

هر یک از این موارد را در ادامه به تفصیل مورد بحث قرار خواهیم داد.

مداخلههای تدریجی به جای مداخلههای یکباره

برخی از مداخلات در یک نقطه زمانی معین شروع می‌شوند، و به سرعت در سراسر جمعیت پخش می‌شوند. مداخله راهنمای تلفن سینسیناتی‌بل نمونه‌ای از این حالت است؛ هزینه پولی، در یک روز معین اعمال شد، و برای همه تلفن‌های راهنما در نظر گرفته شد. اما برخی از مداخلات به تدریج انتشار می‌یابند، به گونه‌ای که ممکن است در این دوره زمانی، رویدادهای دیگری نیز رخ دهند که بر نتایج تأثیر گذاشته، و سبب شوند که گذشت زمان به یک تهدید احتمالی برای روایی درونی تبدیل شود.

هلدر و واگنار (Holder and Wagenaar, 1994) مثالی از مداخله‌ای که به تدریج اجرا می‌شوند، ارائه می‌کنند. آنها آثار آموزش اجباری گارسونها را بر میزان تصادفات جاده‌ای بررسی کردند، و مشخص شد که این طرح، سطح سمّیت خون و رانندگی پرخطر رانندگان مست را کاهش داده است. ایالت اورگان، در دسامبر ۱۹۸۶ این آموزش را پیاده کرد، اما همه افراد نمی‌توانستند یکباره آموزش ببینند. بر اساس تصمیم کمیسیون در اورگان، حدود ۲۰ درصد همه گارسونها تا انتهای ۱۹۸۷ آموزش می‌دیدند، ۴۰٪ تا انتهای ۱۹۸۸ و بیش از ۵۰٪ تا انتهای ۱۹۸۹ آموزش می‌دیدند. اگر اثر روی دریافت‌کنندگان مداخله به سرعت رخ می‌داد، تغییر بلافاصله در سطح کوچک، اما افزایش تدریجی شیب می‌توانست مشهود باشد.

برای تحلیل چنین مواردی، لازم است که درمورد شکل فرایند انتشار اطلاعات داشته باشیم. به عنوان مثال، هلدر و واگنار (Holder and Wagenaar, 1994) به جای متغیر مجازی دومقداره[۳۶۳] $(1,0)$ (که تابع پله-ای نامیده می‌شود، و در آن فرض می‌شود که همه افراد، در همان روز تصویب قانون، تعلیمات را دریافت کرده‌اند)، از گارسونهای تعلیم‌دیده به عنوان متغیر مداخله بهره گرفتند. اگر یک فرایند با پخش کند را به صورت یک تابع تک گام در نظر بگیریم، با مشکلات جدی مواجه خواهیم شد. اول اینکه، تا حد زیادی بر شانس تکیه خواهیم داشت و احتمال بدست‌آوردن نتایج نادرست افزایش می‌یابد، علی‌الخصوص اگر محققان، تابع پله‌ای را به نقطه‌ای در سری زمانی تخصیص دهند که انحراف در نتایج در آن نقطه، در بالاترین حد ممکن باشد. دوم اینکه محقق ممکن است از آثار اولیه کوچک اما واقعی چشم‌پوشی کند، با این فرض که شروع مداخله (۱۹۸۶ برای قانون اورگان) مقارن با نقطه بیشینه تأثیر موردنظر بوده است (نقطه‌ای که در واقع تا حدود سال ۱۹۸۸ حاصل نشد، و تا آن زمان هم تنها ۵۰٪ گارسونها آموزش دیده بودند. از سوی دیگر این شغل، جابجایی شغلی زیادی دارد و بسیار احتمال دارد که افراد آموزش‌دیده شغل خود را نیز تغییر داده باشند). سوم اینکه، حتی زمانی که محققان انتشار مداخله را به دقت مدلسازی کردند، ممکن است به دنبال الگوهایی بگردند که نشان‌دهنده آن انتشار باشد. با این وجود، انتظار وجود

[363] Dichotomous

چنین الگوهایی، ساده‌انگارانه است، زیرا آستانه‌های (حدهای) علّی می‌توانند باعث شوند بروز یک اثر نیازمند رسیدن به سطح معینی از مداخله باشد. هلدر و واگنار (Holder and Wagenaar, 1994) اینطور حدس زدند که آموزش دادن به تنها یکی از چندین گارسون یک رستوران نمی‌تواند روش چندان اثرگذاری باشد (مثلاً اگر دیگر گارسونها به وی فشار بیاورند که مانند قبل سرو کند)؛ بنابراین، باید اکثر گارسونها آموزش ببینند تا فرایند مؤثر باشد. بدون داشتن دانش راجع به نرخ دقیق و شکل انتشار، اغلب بهترین کار این است که به دنبال آثار باتأخیری باشیم که در زمانی بعد از شروع مداخله رخ می‌دهد. بی‌اثر کردن اثر گذشت زمان، یعنی اثراتی که می‌تواند در فاصله زمانی بین شروع مداخله، و زمان بروز تغییر در سری‌زمانی ایجاد شوند (مثلاً تغییر در اجرای قوانین رانندگی)، کار دشواریست.

در این رابطه، بد نیست در نظر بگیریم که چرا داده‌های مطالعه الکل‌سنجی (شکل۶.۶) تا این اندازه واضح بودند. اگر یک قانون جدید کاملاً عمومیت پیدا نکند، یا به صورت ضعیف اجرا شود، انتظار می‌رود که تنها تأثیر تدریجی در واکنش عمومی داشته باشد، و در نتیجه، در مثال الکل‌سنجی، کاهش شدیدی در تصادفات جاده‌ای مشاهده نشود. خوشبختانه از روی داده‌های پس‌زمینه‌ای می‌دانیم که از همان روزی که طرح الکل‌سنجی بریتانیا توسط پلیس اجرا شد، عمومیت هم پیدا کرد. این کار احتمالاً فرایند معمول اطلاع‌رسانی به عموم راجع به کاربرد الکل‌سنج را تسریع می‌کرد، و همچنین بر بکارگیری این دستگاه توسط پلیس در بازه‌های زمانی مکرر بعد از شروع بکارگیری آن نیز تأثیر داشت. در این شرایط، فرایند انتشار واقع‌شده را با تابع پله‌ای بهتر می‌توان نشان داد.

علّیت تأخیردار

همه آثار لحظه‌ای نیستند. حتی زمانی که مداخله به سرعت اجرا می‌شود، ممکن است علّیت‌های تأخیرداری مشاهده شوند؛ مثل آثار باتأخیر سیگار کشیدن بر سرطان ریه. تا دهه‌ها بعد از شروع مصرف سیگار، سرطان ایجاد نمی شود. اگر اجرای مداخله هم به صورت تدریجی باشد، ممکن است باز هم تأخیر بیشتری مشاهده شود. به عنوان مثال، شکل ۶.۳ یک اثر دارای تأخیر را نشان می‌دهد، زیرا برای تولید بطری‌هایی برچسب هشدار داشته باشند، و سپس راه‌یابی این بطری‌ها به قفسه مغازه‌ها، و در نهایت رسیدن به دست مشتری، زمان لازم است. اگر یک نظریه مبنای قوی وجود داشته باشد که به کمک آن بتوانیم تأخیر را پیش‌بینی کنیم (مثل تأخیر ۹ ماهه بین انعقاد نطفه و تولّد که کمک می‌کند تا دریابیم که برچسب‌های هشدار الکل اولین تأثیر خود روی نوزادان را چه زمانی نشان می‌دهند)، علّیت تأخیردار مشکل خاصی ایجاد نخواهد کرد. با این وجود، در بسیاری از اوقات چنین نظریه‌ای وجود ندارد، و در نتیجه، تفسیر یک اثر دارای تأخیر به علت رویدادهای زمانی که بین شروع اجرای مداخله، و شروع آثار باتأخیر رخ می‌دهند، ناممکن می‌شود. در این موارد محقق باید از یک طرح تکرار جابجاشونده بهره گیرد تا بتوان بررسی کرد که آیا با تکرار طرح، همان بازه‌های تأخیر بین شروع مداخله و مشاهده اثر دیده می‌شود؟ این کار تهدید گذشت زمان را کاهش می‌دهد. با این وجود، در این رویکرد فرض می‌شود که مداخله با

واحدهای مختلفی که در زمان‌های مختلف مداخله را دریافت می‌کنند، یا با لحظات تاریخی متفاوتی که هر گروه، مداخله را در آن لحظه تجربه می‌کند، برهم‌کنشی ندارد. به عنوان مثال، افراد دائم‌الخمر ممکن است به برنامه رانندگی در مستی کندتر از دیگر مصرف‌کنندگان اجتماعی پاسخ دهند؛ یا فراگیربودن رسانه‌هایی که یک تصادف ناشی از رانندگی در مستی (DWI) را در یک شهر گزارش می‌کنند، می‌تواند به تأثیر بهتر قانون DWI در آن شهر، در مقایسه با شهرهای دیگری که در آنها قانون بدون کمک این رسانه‌های فراگیر اجرا می‌شود، کمک کند.

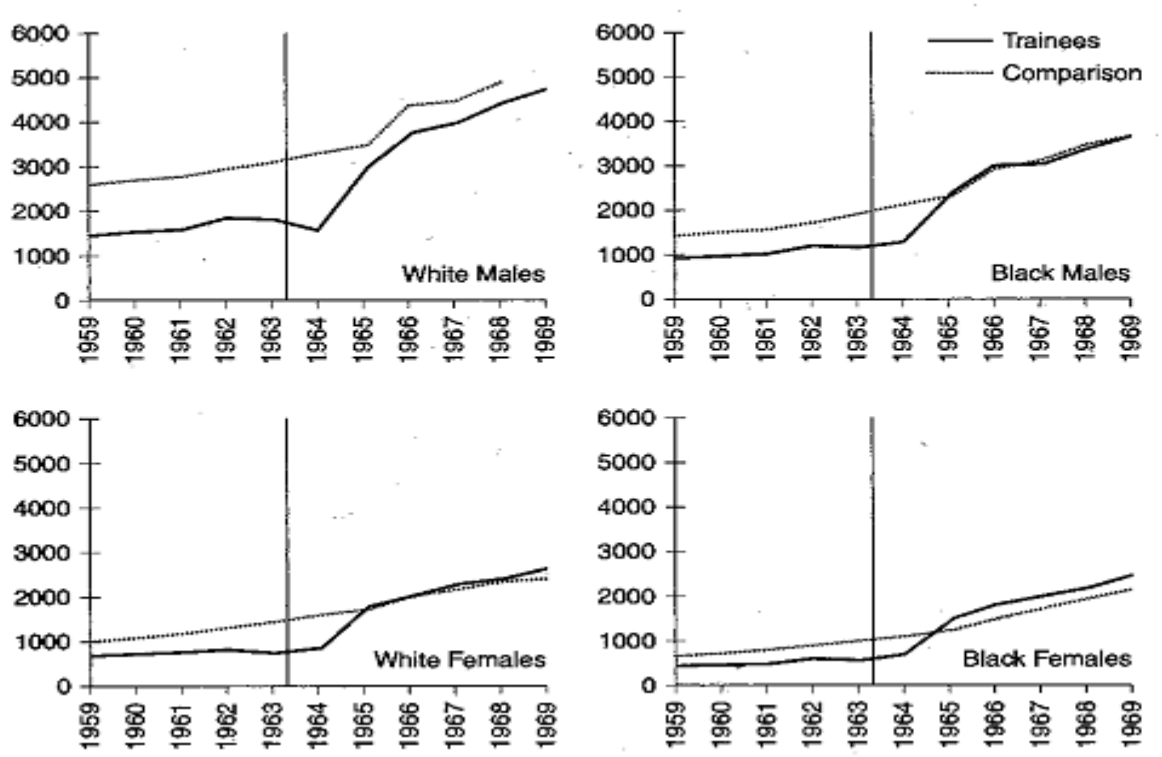
زمانی که آثار باتأخیر همراه با پخش کُند مداخله رخ می‌دهند (همانند شکل ۶.۳)، استنباط علّی بسیار دشوار است. این مسأله بدین دلیل است که هیچ دانشی راجع به این وجود ندارد که مکان شروع اثر موردانتظار یا مطلوب کجا بوده، و می‌توان در هر نقطه‌ای بعد از شروع مداخله، انتظار مشاهده آثار را داشت. هر چه زمان بعد از اجرا طولانی‌تر شوند، تأثیرگذاری عامل‌های مرتبط با گذشت زمان بر تفسیر اثر مداخله تأخیردار بیشتر می‌شود. در این موارد، با افزودن گروه‌های کنترل، متغیرهای وابسته غیرهم‌ارز، حذف مداخله یا تکرارهای جابجاشونده می‌توان استنباطها را بهبود بخشید.

سری زمانی کوتاه

کتابهای موجود در زمینه تحلیل آماری داده‌های سری‌زمانی قوانین متفاوتی برای تعداد نقاط زمانی موردنیاز برای تحلیل مؤثر در نظر می‌گیرند. در بسیاری از آنها پیشنهاد شده است که برای آنکه بتوان روند، تغییرات فصلی، و ساختار خطاهای همبسته در سری را قبل از آزمایش تأثیر مداخله بدست آورد، باید ۱۰۰ مشاهده داشته باشیم. به عنوان مثال، در شکل ۶.۴، نمی‌توان با نگاه کردن به نمودار به راحتی تشخیص داد که یک روند یا الگوی فصلی وجود دارد یا خیر. بررسی بصری زمانی می‌تواند مفیدتر باشد که داده‌های تجمیع‌شده (بازه‌های زمانی) در نمودار ترسیم شده باشند. به عنوان مثال، زمانی که داده‌های شکل ۶.۴ به جای اینکه به صورت هفتگی در نظر گرفته شوند، به صورت فصلی در نظر گرفته شدند، مشخص شد که الگوی تغییر فصلی وجود ندارد؛ اگر چه به نظر می‌رسید که روند کاهشی پایدار مفروض در سری موتورسیکلت‌های کوچک وجود دارد. با این وجود، تجمیع، سبب کوتاه شدن سری‌زمانی می‌شود، که این امر، توانایی مدلسازی دیگر ویژگی‌های داده‌ها را کاهش می‌دهد. بنابراین، بهتر است که در شرایط برابر، تعداد زیادی نقطه داده وجود داشته باشند.

با این وجود، بسیار پیش می‌آید که مشاهدات بسیار بیشتری نسبت به حالتی که تنها یک پیش‌آزمون و پس‌آزمون منفرد موجود است، در دسترس است، اما باز هم تعداد این مشاهدات به ۱۰۰ مشاهده نزدیک نیست. این سری‌های‌زمانی کوتاه، همچنان برای انجام استنباطهای علّی سودمند است، حتی اگر نتوان از تحلیلهای آماری با روشهای استاندارد استفاده کرد. چهار دلیل مهم برای این سودمندی وجود دارد. اول اینکه، افزودن پیش‌آزمون‌های اضافی سبب می‌شود که نسبت به طرح‌هایی که یک یا دو پیش‌آزمون دارند، بهتر بتوان تهدیدهای روایی درونی را بررسی کرد. دوم اینکه، مشاهدات پس‌آزمون اضافی به تعیین مدت تأخیر، و درجه ماندگاری اثر علّی کمک

می‌کنند. سوم اینکه می‌توان از گروه کنترل یا سری‌زمانی کنترل هم بهره گرفت، و این امر، به استنباطهای حاصل از سری‌زمانی کوتاه بسیار قدرت می‌بخشد. چهارم، می‌توان تحلیل‌هایی روی سری‌زمانی کوتاه انجام داد؛ مثلاً به جای توصیف مستقیم خطا، پیشفرضهایی درباره ساختار آن ارائه داد. اقتصاددانان این کار را به طور مکرر انجام می‌دهند (Hsiao, Lahiri, Lee & Pesarsan, 1999 ؛Hsiao, 1986 ؛Greene, 1999).



شکل ۶ – ۱۲. تاثیر شرکت در برنامه آموزش شغلی بر کسب درآمد

Estimating the effects of training programs on earnings, O. Ashenfelter, 1978, Review of Economics and Statistics, 60, pp. 47 – 57. حق تکثیر ۱۹۷۸ برای MIT Press .

سودمندی پیش‌آزمون و پس‌آزمون‌های متعدد

برخی از مزایای سریهای زمانی قطع‌شده کوتاه را می‌توان در شکل ۶.۱۲ مشاهده کرد. این شکل نشان می‌دهد که شرکت در برنامه آموزش ضمن‌خدمت در سال ۱۹۶۴ چه تأثیری روی میزان درآمد گروه‌های مرد و زنان سیاه‌پوست یا سفیدپوست داشته است (Ashenfelter, 1978). گروه مداخله، همه کسانی بودند که در سه ماه اول سال ۱۹۶۴، تحت قانون آموزش و توسعه قدرت انسانی، در این کلاس‌ها شرکت کرده بودند. اشنفلتر پیش‌بینی می کرد که احتمال موفقیت برای این گروه بسیار بالا باشد. گروه کنترل متشکل از ۰.۱٪ نمونه تاریخچه کار پیوسته وزارت

کار بود (این نمونه، نمونه‌ای تصادفی از درآمد ثبت شده کارگران آمریکاییست). متغیر نتیجه‌ای (خروجی) عبارت بود از درآمد افراد، که در ۱۱ نقطه زمانی برای هر یک از چهار گروه اندازه‌گیری شد.

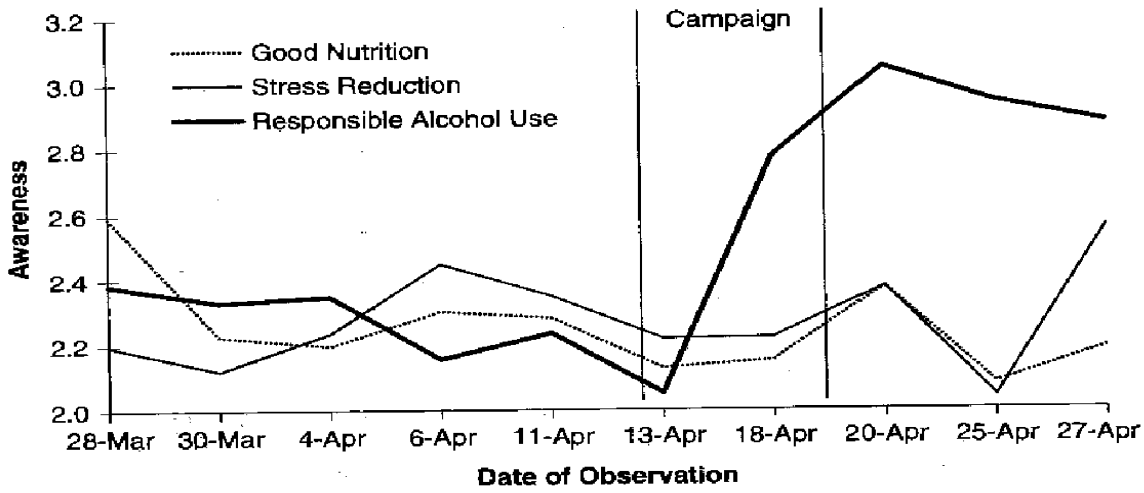
با توجه به اینکه گروه کنترل که در زمره نیروی کار شاغل بودند، درآمد اولیه بالاتری داشتند (داده‌های درآمدی از مقادیر ثبت شده در پایگاه اطلاعاتی تأمین‌اجتماعی گرفته شده‌اند). شکل ۶.۱۲، تأثیر علّی سریع در هر چهار گروه را نشان می‌دهد. حال تصوّر کنید که تنها داده‌های ۱۹۶۳ و ۱۹۶۵ موجود باشند، یعنی اولین سال‌های قبل و بعد از برنامه. باز هم می‌توان افزایش در کسب درآمد را مطرح کرد، تبیینهای جایگزینی وجود دارند که می‌توانند این استنباط علّی را تهدید کنند. یکی از این تبیینها، تهدید بلوغ–انتخاب است؛ یعنی این احتمال که گروه آموزش‌دیده با سرعت بیشتری نسبت به گروه کنترل افزایش درآمد داشته باشد، اما درآمد اولیه پایینتری نسبت به گروه کنترل داشته است (حتی قبل از ۱۹۶۳). با سریهای کوتاه پیش‌آزمونی می‌توان احتمال وجود اختلاف در گروهها به علت بلوغ را به طور مستقیم بررسی کرد.

حال تهدید رگرسیون را در نظر بگیرید. اشخاصی که متقاضی شرکت در برنامه آموزش شغلی در سال ۱۹۶۴ بودند، کسانی بودند که در سال ۱۹۶۳ بیکار بودند. این واقعیت می‌توانسته برآوردهای درآمد سال ۱۹۶۳ کارآموزان را به نسبت سال‌های پیشتر که شاغل بوده، و درآمدی مشابه با درآمد گروه کنترل در سال ۱۹۶۳ داشته‌اند، کاهش داده باشد. اگر چنین شرایطی وجود داشته، و اگر بیکاری این افراد موقتی بود باشد، درآن صورت درآمد گروه مداخله در شرایط پس‌آزمون در هر حال افزایش می‌یافت. اگر تنها اطلاعات سال ۱۹۶۳ را به عنوان تنها نقطه پیش‌آزمون در اختیار داشته باشیم، نمی‌توانیم احتمال وجود تهدید رگرسیون را ارزیابی کنیم، اما با داشتن داده‌های مربوط به چند سال قبل، این کار امکان‌پذیر خواهد شد. در این حالت، رگرسیون می‌توانسته به اثر آموزش اضافه شده باشد، زیرا بین سال‌های ۱۹۶۲ و ۱۹۶۳، کاهش اندکی در درآمدهای گروه مداخله مشاهده شد. اما رگرسیون نمی‌توانست دلیل همه اثرات مداخله باشد، زیرا میانگین درآمد پیش‌آزمون گروه مداخله بسیار کمتر از مقدار درآمد پیش‌آزمون گروه کنترل در طول آن سالها -و نه فقط سال ۱۹۶۳- بود.

بدون داشتن اطلاعات مربوط به سال‌های پس‌آزمون در شکل ۶.۱۲، نمی‌توان به قطعیت گفت که آیا اثر آموزش ماندگاری داشته، یا به سرعت از بین رفته است. بدون داشتن مجموعه‌های پیش‌آزمون ممکن است این سوال مطرح شود که- با در نظر گرفتن تنها داده‌های سالهای ۱۹۶۲ تا ۱۹۶۵- آیا تغییرات مشاهده شده در درآمد در فاصله سالهای ۱۹۶۲ تا ۱۹۶۵، نشان دهنده یک دوره چهار ساله چرخه اقتصادی در یک روند کلی رو به بالا بوده است؟ از آنجا که در سال‌های ۱۹۵۹ تا ۱۹۶۲ هیچ طرح چرخه‌ای مشاهده‌نشده، بنابراین می‌توان این احتمال را نادیده گرفت. بنابراین، افزودن پیش‌آزمون‌ها و پس‌آزمون‌های مختلف می‌تواند کمک قابل توجهی به تفسیر بهتر شبه‌آزمایش‌ها کند، حتی اگر انجام سری‌زمانی کامل امکان‌پذیر نباشد (H. Bloom, 1984b این داده ها را مجدداً مورد تحلیل قرار داده است).

تقویت سریهای کوتاه با افزودن عناصر طراحی

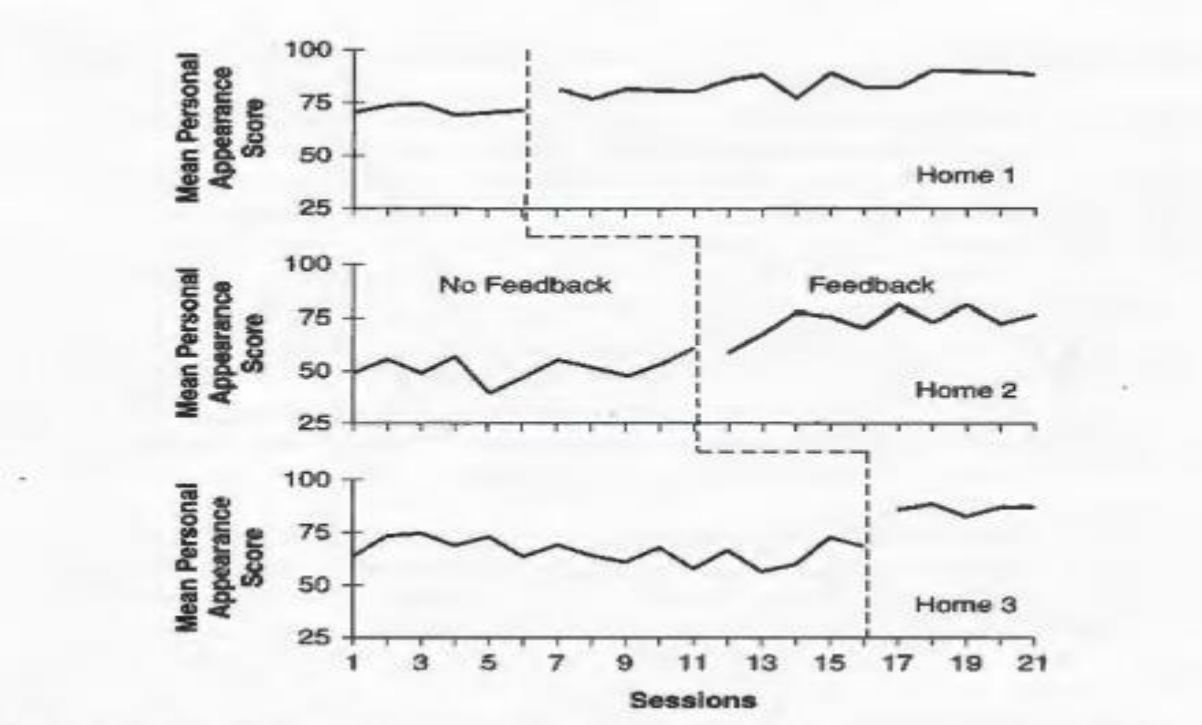
قابلیت تفسیرپذیری سری‌زمانی کوتاه با افزودن هرکدام از عناصر طراحی که در این فصل یا فصل‌های پیشین (جدول ۵.۲) مطرح شد (مانند گروه کنترل، متغیرهای وابسته غیرهم‌ارز، تکرار جابجاشونده، حذف مداخله و تکرارهای متعدد) بهبود می‌یابد (مثلاً Barlow & Hersen, 1984؛ R. Franklin, Allison & Gorman, 1997؛ Sidman, 1960؛ Kratochwill & Levin, 1992). به عنوان مثال، مک‌کیلیپ (McKillip, 1992) آثار یک کمپین رسانه‌ای در سال ۱۹۸۹ را در کاهش استفاده از الکل در فستیوال‌های دانشجویی در کمپ‌های دانشگاه ارزیابی کرد. متغیر وابسته اولیه او، سری‌زمانی کوتاهی (۱۰ مشاهده) در مورد آگاهی نسبت به مصرف نادرست الکل در جمعیت هدف بود. برای تقویت استنباط در این سری‌کوتاه، مک‌کیلیپ دو متغیر وابسته غیرهم‌ارز که از نظر مفهومی با سلامت ارتباط داشتند را به مدل افزود (او آنها را سازه‌های کنترل نامید، تا بر شباهت آنها به گروه کنترل تأکید کرده باشد). اگر اثرات به دلیل ارتقاء کلی نگرش نسبت به سلامت بود، این دو متغیر تغییر می‌کردند. اما این دو متغیر (تغذیه خوب، کاهش استرس) هدف کمپین نبودند، بنابراین اگر اثرات مشاهده شده بدلیل مداخله ایجاد شده باشند، این دو متغیر نباید تغییر کنند. همانگونه که شکل ۶.۱۳ نشان می‌دهد، آگاهی از سوءمصرف الکل مشخصاً در طول کمپین افزایش یافته است اما آگاهی از دیگر مسائل مرتبط با سلامتی افزایش نداشته است (برای مشاهده دیگر کاربردهای این تحقیق که موجب کسب نتایج جاه‌طلبانه‌تری نیز شد نگاه کنید به، Fischer, 1994).



شکل ۶ – ۱۳. اثر یک برنامه تلویزیونیp بر روی افزایش آگاهی نسبت به سوء مصرف الکل

از Research without control groups: A control construct design: by J. McKillip, 1992, Methodological issues in applied psychology, edited by F. B. Bryant, J. Edwards, R. S. Tindale, E.J. Posavac, L. Heath & E. Henderson, New York: Plenum. حق تکثیر ۱۹۹۲ Plenum Press.

مک‌کلاناهان، مک‌گی، مک‌داف و کرانتز (McClannahan, McGee, MacDuff and Krantz, 1990) عنصر تکرارجابجاشونده را به سری‌زمانی کوتاه (۲۱ مشاهده) خود افزودند. مطالعه آنها اثرات دادن بازخورد منظم به زوج‌های متاهلی که سرپرستی خانه‌های گروهی کودکان اوتیست را بر عهده داشتند، را بر بهداشت فردی روزانه و ظاهر کودکان در این خانه‌ها بررسی می‌کرد. در خانه شماره ۱ بازخورد بعد از جلسه ۶ ام داده شد، در خانه شماره ۲ بعد از جلسه ۱۱ام، و در خانه ۳ بعد از جلسه ۱۶ ام. بعد از هر بازخورد، ظاهر و بهداشت شخصی کودکان در آن خانه به بالاتر از خط پایه افزایش یافته، و این بهبود پیوسته در طول زمان حفظ می‌شد (شکل۶.۱۴). با این وجود، هر دو این مثال‌ها، ایراد سری‌زمانی کوتاه، یعنی دشواری در تعیین مدت زمان باقی ماندن اثر، را نشان می‌دهند. شکل ۶.۱۳ این مسأله را به وضوح نشان می‌دهد. در این نمودار، آغاز کاهش در آگاهی نسبت به سوء مصرف الکل، بعد از دو هفته اجرای مداخله، به خوبی مشهود است.



شکل ۶ – ۱۴. اثر مداخلات والدین بر ظاهر فیزیکی کودکان اوتیستی در سه خانه متفاوت.

از Assessing and improving child care: A personal appearance index for children with autism" by L. E. McClannahan et al., 1990. Journal of applied behavior analysis, 23, 469 – 482. Copyright 1990 by the society for the experimental analysis of behavior.

تحلیل سری‌زمانی کوتاه

بسیاری از محققین قادر نیستند داده‌های بدست‌آمده از سریهای‌زمانی را تحلیل کنند. برخی بر این باورند که بررسی بصری کافی است؛ و اگر اثری آنقدر کوچک باشد که برای شناسایی آن نیاز به بررسیهای بیشتر آماری باشد، این اثر ارزش جستجو کردن ندارد. اما تحقیقات نشان می‌دهد که در بررسی بصری آثار کوچک یا باتآخیر، فراوان امکان اشتباه وجود دارد (مثلاً Furlong & Wampold, 1981؛ Ottenbacher, 1986؛ Wampold & Furlong, 1981) و با توجه به وجود آثار کوچکی در گذشته که اهمیت فراوانی داشته اند (مثلاً Rosenthal, 1994)، این باور که اثرات کوچک اهمیتی ندارند (مثلاً Barlow & Hersen, 1984 صفحه ۲۸۲) نیز باید مورد بازنگری قرار گیرد. دیگر محققین- با نظرات متفاوتی در مورد ارزش یافتن این اثرات- گزینه‌های تحلیلی مختلفی را [برای تحلیل سری‌های زمانی کوتاه] مورد بررسی قرار داده‌اند (مثلاً Franklin, Allison, & Gorman, 1997؛ B. Gorman & Allison, 1997؛ Matyas & Greenwood, 1997). از جمله گزینه‌های ناپارامتری بالقوه صحیح می‌توان به آزمایش‌های (دقیق) تصادفی‌سازی (Edgington, 1992؛ Gorman & Allison, 1997؛ Koehler & Levin, 1998) و روش‌های بوت‌استرپ[۳۶۴](Efron & Tibshirani, 1993) اشاره داشت؛ اگرچه هر دو روش توان اندکی در تحلیل در سری‌های زمانی کوتاه دارند. کی. جونز (K. Jones, 1991) و کروزبی (Crosbie, 1993) روش‌های پارامتریک جایگزین را پیشنهاد می‌کنند، اگرچه نقدهای جدی به مشکلات قابل‌توجه موجود در این روشها وارد است (Reichardt, 1991). اگر اندازه‌گیری در طول زمان، روی افراد یکسانی انجام شود، می‌توان از انواع مختلفی از ANOVAها با اندازه‌گیری مکرر، مدل‌های منحنی رشد یا تحلیل تاریخی رویداد استفاده کرد (مثلاً Carbonari, Wirtz, Muenz & Stout, 1994). اقتصاددانان نیز هنگام کار با سری‌زمانی کوتاه، فرض‌هایی را در مورد ساختار خطای موجود در داده‌ها تدوین می‌کنند تا بتوانند برآوردهای خطای استاندارد را تصحیح نمایند (Greene, 1999؛ Hsiao, 1986, 1999). اگر بتوان یک سری‌زمانی کوتاه را از افراد متعددی بدست آورد، تجمیع[۳۶۵] سریهای‌زمانی می‌تواند سودمند باشد (Sayrs, 1989؛ West & Hepworth, 1991). شاید بهترین توصیه، بکارگیری چندین روش آماری برای تحلیل سری‌های زمان کوتاه باشد (برای دیدن خلاصه‌ای از این گزینه‌ها به مقاله Allison & Gorman, 1997 مراجعه کنید). اگر نتایج تحلیل‌های مختلف همگرا بوده، و با بررسی‌های بصری نیز سازگار باشند، اطمینان به یافته‌های تحلیلها به نحو بهتری تضمین می‌شود.

با این وجود، تسهیل تحلیل بصری داده‌های سری‌زمانی اهمیت ویژه‌ای دارد. نمایش بصری خوب با قدرت گرافیکی بالا، خواننده را به ذوق می‌آورد. مطالعه تکنیک‌های گرافیکی خود یک تخصص است (مثلاً Tufte, 1983, 1990)، و می‌توان به کمک رایانه، طیف گسترده‌ای از تکنیک‌های گرافیکی با تنوع روزافزون را در اختیار داشت. تحلیل

[364] Bootstraping
[365] Pooling

آماری و تحلیل بصری هردو (با در نظر گرفتن محدودیتهای هر کدام از آنها) در تحلیل سریهای‌زمانی حائذ اهمیت هستند.

محدودیتهای موجود در بسیاری از داده‌های آرشیوی

بسیاری از داده‌های مورد استفاده در سری‌زمانی از آرشیوهایی که در موسسات خصوصی و دولتی نگهداری می‌شوند بدست می‌آیند. با این وجود، معمولاً به سختی می‌توان از وجود این آرشیوها اطلاع کسب کرد و در صورت شناسایی آنها نیز، دسترسی پیدا کردن به داده‌های بخش تجاری خصوصی یا موسسات محلی (مانند مدارس یا شهرداری‌ها) غالباً دشوار است. اگرچه در سال‌های اخیر، بهبودهایی در این زمینه حاصل شده. به عنوان مثال، کیکولت و ناتان (Kiecolt and Nathan, 1990) برخی از خدمات آرشیوی بسیار پیچیده که در ایالات متحده آمریکا وجود دارد را شرح می‌دهند. شبکه‌های رایانه‌ای مختلف سبب شده‌اند که تعیین مکان و بازیابی این داده‌های سری‌زمانی، بسیار راحت‌تر شود.[۳۶۶]

مشکلات مربوط به روایی سازه در داده‌های آرشیوی نیز اهمیت فراوانی دارند. از آنجا که اغلب داده‌های گردآوری و ذخیره شده، برای پایشهای (مانیتورینگ) اجتماعی و اقتصادی به کار گرفته می‌شوند، متغیرهایی که به جای «خروجی یا نتیجه» رنگ و بوی «فرایند» دارند، مورد توجه بیشتر هستند. همچنین مقادیر آرشیوی مستقیم در مورد فرایندهای روانشناختی و یا گروه‌های کوچک به ندرت یافت می‌شوند. در نتیجه، آرشیوهای کنونی ممکن است برای بررسی توضیحات علّی یا ساختارهای فیزیولوژیکی مناسب نباشند[۳۶۷].

استفاده از انواع مختلف داده‌های آرشیوی نیازمند بررسیهای دقیق است. تعاریف عملیاتی باید به دقت بررسی شوند؛ ممکن است عناوین سازه‌های بکاررفته در یک اندازه‌گیری خاص بخوبی انتخاب نشده باشند. به همین ترتیب، باید تغییرات رخ داده در تعاریف، در طول زمان را نیز مورد بررسی قرار داد. اگر امکان داشته باشد، ماهیت این تغییرات باید مستند شود. اینکار بسیار کمک‌کننده خواهد بود علی‌الخصوص اگر داده‌های هم‌پوشان برای دوره‌های خاصی که در آن هم از تعاریف قدیم و هم از تعاریف جدید استفاده شده است، در دسترس باشد. در بسیای از موارد، داده‌های برخی از زمان‌ها از دست رفته است، و اگر این مشکل خیلی جدی نباشد، می‌توان از آن صرف‌نظر کرد. اگر داده‌های رسم شده، نظم مشکوکی داشته باشند، مثلاً مقادیر در یک بخش سری ثابت باقی بمانند، یا افزایش در واحد زمان ثابت باشد، ممکن است که داده‌ها به شکل مناسبی گردآوری نشده باشند، یا اینکه فرد برای مشاهدات از دست رفته، مقادیر درونیابی شده ارائه کرده است بدون اینکه به آن اشاره کرده باشد.

---

[۳۶۶] فهرست جامعی از سری‌های زمانی اقتصادی و اجتماعی در http://www.economagic.com/ موجود است و برخی از سری‌های زمانی دولتی ایالات متحده در سایت‌های http://www.fedstats.gov/ و http://www.census.gov/ ارائه شده است.

[۳۶۷] توضیح مترجم: به نظر می رسد بتوان ادعا کرد این مساله با در نظر گرفتن گسترش روزافزون پایگاههای داده ای که اطلاعات مربوط به کنشهای افراد را ذخیره می کنند و همچنین شیوه های پیچیده داده کاوی که امکان تحلیل این کنشها را فراهم می آورند تا حدید زیادی برطرف شده است.

بالعکس، ممکن است داده‌ها بسیار گسترده باشند، و در مقابل هر گونه بازمقیاس برای کاهش تغییرپذیری مقاومت کنند، و شاید دلیلش این باشد که داده‌ها به جای اینکه پدیده‌های ذاتاً ناپایداری باشند، و به صورت نامنظم گردآوری شده‌اند.

مشکل اصلی داده‌های آرشیوی، انعطاف‌ناپذیری آنهاست. تحلیل‌گران سریهای‌زمانی، داده‌هایی را ترجیح می‌دهند که بتوان آنها را به بازه‌های زمانی پرفراوانی‌تر، نواحی محلی، دموگرافیهای فردی، و زیربخشهای موضوعی جزئی تجزیه کرد (Campbell, 1976). اما آرشیوها معمولاً این انعطاف‌پذیری را ندارند. اغلب، محققان به دنبال داده‌های هفتگی، ماهانه و فصلی هستند، زیرا سریهای طولانی‌تری نسبت به داده‌های سالانه تولید می‌کنند، حساسیت بیشتری نسبت به اثرات علّی بلافاصله دارند، و بهتر می‌توان با استفاده از آنها تبیینهای جایگزین مرتبط با گذشت زمان را بی‌اثر نمود. اما اگر داده‌ها به شکل سالانه گردآوری و ذخیره شده باشند، تجزیه آنها به صورت واحدهای زمانی کوچکتر ممکن نیست. محققان معمولاً می‌خواهند که داده‌ها را بر مبنای متغیرهای مختلف اجتماعی مانند مشخصات دموگرافیک از جمله نژاد، طبقه یا جنسیت، تقسیم‌بندی کنند. این تجزیه بررسی چگونگی تغییر اثرات در گروه‌ها یا بررسی روایی بیرونی یا گروه‌های دیگری که مداخله دریافت نکرده و گروه کنترل را شکل می‌دهند، را ممکن می‌سازد. به عنوان مثال، در مطالعه آثار الکل‌سنجی، اگر بتوان داده‌ها را به الکلی‌ها و غیرالکلی‌ها تقسیم-بندی کرد، یا به صورت گروه‌های مذهبی مختلفی که نوشیدن الکل را مجاز می‌دانند یا نمی‌دانند تقسیم‌بندی کرد، نتایج مفید خواهند بود. اغلب وجود متغیرهای وابسته بیشتر به محقق کمک میکند، زیرا محقق می‌تواند یک متغیر وابسته غیرهم‌ارز در میان آنها پیدا کند. با این وجود، در تمامی این موارد، محققان باید به هر آنچه چه که در آرشیو موجود است، قانع باشند.

البته نباید بیش از اندازه نسبت به صلبیت یا کیفیت داده‌های آرشیوی بدبین باشیم. همانگونه که مثال‌های این فصل نشان می‌دهند، می‌توان طرحهای سری‌زمانی قطع‌شده با قابلیت تولید استنباطهای علّی مطمئنی را نیز اجرا نمود. به علاوه، با کمی پشتکار و هوشیاری، محقق می‌تواند داده‌هایی را در آرشیوها بیاید که انتظار پیدا کردنشان را ندارد. به عنوان مثال، کوک، کالدر و وارتون (Cook, Calder and Wharton, 1979) یک سری ۲۵ ساله را پیدا کردند که طیف وسیعی از متغیرها را پوشش می‌داد، و امکان طبقه‌بندی آنها بر اساس مصرف، لذت، رفتار سیاسی، مشارکت نیروی کار، ساختار اقتصادی محلی، سلامت عمومی و جرم و جنایت وجود داشت. اگرچه برخی از این داده‌ها از آرشیوهای فدرال گرفته شده بودند، اما اغلب آنها از داده‌های ایالتی مهجوری بودند که به نظر نمی‌رسید با هدف تحقیقات جمع‌آوری شده باشند. ایالت‌ها از نظر کیفیت داده‌هایی که نگهداری می‌کنند با یکدیگر متفاوت هستند. برای اینکه یک ناحیه مهم به طور کافی پوشش کافی داده شود، ممکن است لازم باشد محققان برخی متغیرها را در یک ایالت، و برخی دیگر را در ایالاتی دیگر بیابند. با این وجود، مقادیر شگفت‌آوری از داده‌های سری‌زمانی وجود دارند. اگر داده‌های جدید و آینده، کیفیت فنی بهتری نسبت به داده‌های جمع‌آوری شده در

گذشته داشته باشند، می‌توانیم امیدوار باشیم که در آینده از این منابع به میزان بیشتری در تحلیل‌های سری‌زمانی بهره گرفته شود.

نکاتی در باب سری‌های زمانی ملازم یا همزمان[۳۶۸]

طرح‌های سری‌زمانی قطع‌شده که در فصل حاضر مورد بررسی قرار گرفت، با دیگر کاربرد سریهای‌زمانی که گاه با هدف استنباط علّی مورد استفاده قرار می‌گیرد - یعنی سری‌زمانی همزمان- تفاوت اساسی دارد. سری‌زمانی قطع‌شده، به یک مداخله نیاز دارد که با دقت اجرا و مدیریت شود. گاهی ممکن است یک عامل علّی بالقوه، به این روش بکار گرفته نشود، بلکه شدت آن - بدون وجود گروه کنترل- در یک دوره زمانی که در طول آن متغیر خروجی نیز تغییر می‌کند، نوسان داشته باشد. در یک سری زمانی همزمان، محقق میان سری‌زمانی بدست آمده از متغیر علّی پیش‌فرض و سری‌زمانی متغیر نتیجه‌ای (خروجی) پیش‌فرض همبستگی برقرار می‌کند؛ هر دو سری در طول یک بازه زمانی مشترک، و در مورد واحدهای (افراد) مشابه اندازه‌گیری می‌شوند. سپس محققان نحوه ارتباط فراز و فرودها در سری علّی با فراز و فرودهای متعاقب (از نظر زمانی) در سری متغیر خروجی را مورد بررسی قرار می‌دهند. این کار به مسأله تقدم زمانی در مفهوم‌پردازی علّیت، محوریت می‌بخشد. مک‌کلیری و ولش (McCleary and Welsh, 1992) به مطالعه‌ای منتشرشده درباره همبستگی میان ارجاعات به تست شخصیت چندمحوری مینه‌سوتا[۳۶۹] (MMPI) و تست رورشاخ[۳۷۰] اشاره می‌کنند. این مطالعه به این فرضیه می‌پرداخت که هر چه تعداد بیشتری از پزشکان از تست شخصیت MMPI بهره گیرند، این تست بیشتر و بیشتر جای رورشاخ را به عنوان یک گزینه تست شخصیت می‌گیرد. همبستگی‌های گزارش‌شده در مطالعه این فرضیه را تأیید می‌کنند، اگرچه تحلیل اصلی، مشکلات آماری فراوانی داشت؛ از قبیل اینکه خودهمبستگیها[۳۷۱] در نظر گرفته نشده بودند، و مطالعه نتوانسته بود از طریق محاسبه همبستگی‌های تأخیردار، تأخیرهای زمانی را لحاظ نماید.

نکته مهم این است که در سری زمانی هم‌زمان، علّت مفروض به صورت آزمایشی دستکاری نمی شود، بلکه به شیوه‌ای کنترل‌نشده نوسان می‌کند. بنابراین با توجه به دلایل متقن موجود برای این مسأله که همبستگی اثبات- کننده علّیت نیست، ممکن است اینکه همبستگی‌های کنترل‌نشده شاهدی برای علّیت در نظر گرفته شوند، عجیب به نظر برسد. برخی از طرفداران این روش با اتکاء به منطق گرنجر (Granger, 1969)، نام «علّیت گرنجر[۳۷۲]» را به این نوع همبستگیها اطلاق می‌کنند. منطق گرنجر عبارتست از اینکه اگر رابطه علّی در یک دوره تأخیری خاص، تک‌جهته باشد، و اگر دو متغیر شرایط تحلیل برای «مزاحم سفید[۳۷۳]» را برآورده سازند، همبستگی (با تأخیر زمانی

[368] Concomitant time series
[369] Minnesota Multiphasic personality inventory
[370] Rorschach
[371] Autocorrelation
[372] Granger causality
[373] White noise

متناسب) برآوردی بدون‌سوگیری از رابطه علّی بدست می‌دهد. متأسفانه، با وجود آنکه می‌توان شرط «مزاحم سفید» را آزمون کرد، اما امکان برقراری شرط‌های دیگر در عمل وجود ندارد. برقراری شرط علّیت تک جهته ساده در کاربردهای دنیای واقعی بسیار نامحتمل است (McCleary & Welsh, 1992)؛ شونکوف و فیلیپس (Shonkoff and Phillips, 2000) بیان کردند که این مسأله، اغلب سوگیری همزمانی[۳۷۴] نامیده می‌شود. کرومول، هانان، لابیز و ترزا (Cromwell, Hannan, Labys and Terraza, 1994) نیز چنین نتیجه‌گیری می‌کنند که «وقتی درباره علّیت گرنجر صحبت می‌کنیم، در واقع بررسی می‌کنیم که آیا یک متغیر خاص مقدم بر دیگری است یا خیر، و علّیت را از نظر رابطه علّت و معلول بررسی نمی‌کنیم» (صفحه ۳۳، Holland, 1986، Menard, 1991؛ Reichardt, 1991 را نیز ببینید)[۳۷۵].

نتیجه‌گیری

سری زمانی قطع‌شده، یکی از انواع مهم طرحها برای بدست‌آوردن دانش علّی است، در شرایطی که بتوان داده‌های موردنیاز را از نقاط زمانی متعدد گردآوری کرد. مزیت این طرحها در این است که در این سری‌ها، وجود اندازه‌گیریهای پیش‌آزمونی متعدد، امکان بررسی احتمال وجود بسیاری از تهدیدها را فراهم می‌آورد. اطلاع دقیق از زمان اجرای مداخله کمک می‌کند تا بتوان با تهدید گذشت زمان مقابله کرد، و داده‌های پس‌آزمونی سبب می‌شوند تا بتوان شکل رابطه علّی را بر حسب سرعت شروع و درجه پایداری اثر توصیف کرد. نگارندگان این کتاب در زمره طرفداران دوآتشه طرح‌های سری‌زمانی قطع‌شده هستند، و استفاده از طرحها را - خواه با استفاده از داده‌های آرشیوی، و خواه با داده‌های دست‌اول جمع‌آوری‌شده توسط محقق- به پژوهشگران توصیه می‌کنند. حتی زمانی که تعداد مشاهدات کمتر از مقدار موردنیاز برای تحلیل آماری مرسوم هست، باز هم بکرگیری این روش توصیه می شود. اصل این است که هر چه اطلاعات بیشتری درباره ابعاد زمانی پیش‌مداخله و پس‌مداخله عملکرد داشته باشیم، بهتر می‌توانیم ابهام موجود درباره اینکه آیا رابطه همبستگی مدنظر علّیست یا نه را کاهش دهیم.

این طرحها از نظر دانشی که درباره مداخله موردمطالعه بدست می‌دهند نیز دارای مزیت‌اند؛ از جمله زمان شروع و شکل پخش اثر مداخله در جمعیت موردمطالعه. دانستن اینکه مداخله دقیقاً در کدام نقطه از زمان شروع به تأثیرگذاری روی متغیر نتیجه‌ای کرده است، یکی از بهترین مزیت‌های این روش است، به ویژه زمانی که مداخله در یک نقطه خاص در طول یک مقیاس فاصله‌ای با سطوح متعدد رخ می‌دهد. در مورد سری‌زمانی، زمان همان مقیاس است، و قطعی یا شکست در گراف شکلی از پاسخ در زمان مداخله است که علائمی برای تشخیصی علّیت در اختیار می‌گذارد.

---

[374] Simultaneity bias

[۳۷۵] مشکلات مشابهی در مدل‌های پیشنهادی دیگر که علیت را از مشاهدات کنترل نشده روابط بین دو متغیر بدست می‌آورند، وجود دارد (مثلاً Wampold, 1992).

در فصل بعد، بسیاری از این اصول را دوباره به شکلی مشابه مشاهده خواهیم کرد. طرح ناپیوستگی رگرسیونی نیز به اطلاعات مربوط به زمان شروع مداخله به صورت یک نقطه از پیوستار نیاز دارد (اگرچه پیوستار در اینجا زمان نیست، بلکه ترتیب نقاط قرار گرفته بر مقیاس متغیر بکار گرفته‌شده برای تخصیص واحدها به شرایط آزمون و کنترل است)، و اثر مداخله با تغییر شدید، متغیر خروجی در آن نقطه نشان داده می‌شود. طرح ناپیوستگی رگرسیونی، مشخصات مشترکی با طرح‌های آزمایشهای تصادفی که در ادامه این کتاب بررسی خواهیم کرد، دارد. بنابراین به ما کمک می‌کند تا میان سری‌زمانی قطع‌شده و آزمایش تصادفی ارتباط برقرار کنیم.

فصل ۷

طرح‌های ناپیوستگی رگرسیون

ناپیوستگی[۳۷۶]: ۱. فقدان پیوستگی، توالی منطقی یا چسبندگی. ۲. یک جدایی یا شکاف. ۳. در زمین‌شناسی: سطحی که سرعت امواج لرزه‌ای در آن تغییر می‌کند. ۴. دگرگونی. الف. نقطه‌ای که تابع در آن تعریف شده است اما پیوسته نیست. ب. نقطه‌ای که تابع در آن تعریف شده نیست.

وقتی زندانیان از زندان آزاد می‌شوند، فاقد شغل یا منابع مالی هستند، بنابراین نمی‌توانند اعضای مفیدی برای جامعه باشند. آیا برخی از آنها بعد از ترک زندان دوباره جرم خود را تکرار می‌کنند تا منابع مالی بدست آورند؟ آیا ارایه کمک مالی به آنها سبب می‌شود که احتمال ارتکاب تخلف در آینده کاهش یابد؟ برک و رائوما (Berk and Rauma, 1983؛ Rauma & Berk, 1987) سعی کردند تا به این سؤال آخر پاسخ دهند. ایده این مطالعه زمانی به ذهن این دو محقق رسید که ایالت کالیفرنیا قانونی را تصویب کرد که به موجب آن به زندانیان تازه آزاد شده از زندان کمک هزینه بیکاری پرداخت می‌شد، البته با این شرط که زندانیان در ۱۲ ماه قبل از آزادی، در درون زندان ۶۵۲ ساعت کار کرده باشند. کسانی که ساعات کمتری کار کرده بودند، مشمول پرداخت کمک مالی نمی شدند. برک و رائوما دریافتند که کسانی که کمک هزینه بیکاری دریافت کرده بودند، ۱۳ درصد کمتر از گروه کنترل، مجدداً مرتکب جرم شدند. به دلایلی که بیان خواهیم کرد، این برآورد با لحاظ کردن بعضی پیشفرضها، از نظر آماری بدون سوگیری محسوب می‌شود.

[376] Discontinuity

کسانی که فصول پیشین مربوط به شبه‌آزمایش‌ها را مطالعه کرده‌اند، ممکن است تعجب کنند که چگونه مطالعه‌ای با این حجم قابل‌توجه و مشهود از «مشکلات» انتخاب، به استنباط علّی بدون سوگیری منتهی می‌شود. زندانیان آزاد شده به‌خاطر تفاوتی که داشته‌اند، به گروه‌های کنترل و مداخله تخصیص داده شده‌اند، نه به‌خاطر شباهت‌شان (یگ گروه بیشتر از مقدار برش ۶۵۲ ساعت، و دیگری کمتر از این میزان کار کرده بودند). در این فصل نشان خواهیم داد که چرا این طرح (طرح ناپیوستگی رگرسیونی[۳۷۷] (RD)) به صورت گزینشی انجام می‌شود، اما برآوردهای علّی بدون‌سوگیری ارائه می‌دهد.

کار روی طرح RD در سال ۱۹۵۸ آغاز شد (Campbell, 1984) و تیستلویت و کمپبل (Thistlewaite and Campbell, 1960) برای اولین بار این طرح را در مطالعه خود بکار گرفتند. این طرح در مطالعات دیگر حوزه‌ها اعم از پزشکی و بهداشت عمومی (Finkelstein, Levin & Robbins, 1996a, 1996b)، اقتصاد (Goldberger, 1972a, 1972b)، آموزش (Tallmadge & Horst, 1976؛ Tallmadge & Wood, 1978) و در آمار (Rubin, 1977, 1978) نیز به کار گرفته شده است. گلدبرگر (Goldberger, 1972a, 1982b)، لرد و نویک (Lord and Novick, 1968، صفحات ۱۴۰ – ۱۴۴) و رابین (Rubin, 1977) اثبات‌های آماری دقیق ارائه کردند که نشان می‌داد اگر پیشفرضها برقرار باشند، این طرح (یا طرح‌های مشابه) می توانند برآوردهایی بدون‌سوگیری از اثرات مداخله بدست دهند. خلاصه‌ای از این طرح‌ها توسط نویسندگان مختلف (Huitema378, 1980؛ Judd& Kenny, 1981a؛ Marsh, 1998؛ Mohr, 1988, 1995) و به ویژه تروخیم و همکارانش (مثلاً Cappelleri, 1991؛ Trochim, 1984؛ Trochim & Cappelleri, 1992) ارائه شده است. طرح RD بیش از ۲۰۰ بار در ارزیابی برنامه‌های محلی با بودجه دولتی، موضوع قانون آموزش ابتدایی و متوسطه مصوب سال ۱۹۶۵ (در آمریکا)، مورد استفاده قرار گرفته است (Trochim, 1980). بجز این مورد، تنها تعداد اندکی از مطالعات استفاده از این طرح را گزارش کرده‌اند[۳۷۹]. این عدم‌اقبال احتمالاً به علّت مشکلات عملی همراه با این طرحهاست. مشکلاتی که منجر به ایجاد محدودیت‌هایی برای این طرحها می‌شود. فصل حاضر به بررسی این محدودیتها خواهد پرداخت. با این حال، می‌توان طرح را به شکل گسترده‌تر نسبت به آنچه تا کنون بوده، به کار گرفت، و گاهی آنها را جایگزین شبه‌آزمایش‌های بهتر شناخته‌شده‌تری که از نظر استنباط ضعیف‌تر هستند نمود، یا گاهی آن را به شبه‌آزمایش‌های موجود افزود تا استنباطهای علّی بهبود پیدا کنند، و یا گاهی آنها را با آزمایش‌های

---

[377] Regression discontinuity design

[378] هویتما آن را آزمایش تخصیص اریب نامیده است.

[۳۷۹] Abadzi, 1984, 1985؛ Berk & Deleeuw, 1999؛ Berk & Rauma, 1983؛ Braden & Bryant, 1990؛ Cahan, Linchevski, Ygra & Danziger, 1996؛ Cahan & Davis, 1987؛ Cappelleri & Trochim, 1994؛ Carter, Winkler & Biddle, 1987؛ Cullen et al., 1999؛ Deluse, 1999؛ DiRaddo, 1996، Finkelstein et al., 1996b؛ Havassey, 1988؛ Klein, 1992؛ Lipsey, Cordray & Berger, 1981؛ Mark & Mellor, 1991؛ Rauma & Berk, 1987؛ Robinson, Bradley & Stanley, 1990؛ Robinson & Berger, 1981؛ Mark & Mellor, 1991؛ Rauma & Berk, 1987؛ Robinson, Bradley & Stanley, 1990؛ Robinson & Stanley, 1989؛ A. Ross & Lacey, 1983؛ Seaver & Quarton, 1976؛ Stadrhaus, 1972؛ G. Thomas, 1997؛ Visser & deLeeuw, 1984 را ببینید. مشخص است که برخی از این مطالعات، الزامات کامل طراحی را رعایت نکرده‌اند یا مشکلات عملی شدیدی در تجهیزات دارند؛ اما در به هر حال در این فهرست آمده‌اند تا افراد علاقه‌مند بتوانند با دقت بیشتر آنها را بررسی کنند.

تصادفی ترکیب کرد تا توان و اصول اخلاقی هر دو طرح بهبود پیدا کند. امیدواریم که این فصل بتواند با مشخص کردن مزایا و شرایط تسهیل‌کننده این طرح، اهمیت RD را بیان کرده و شرایط اجرای آن را تسهیل نماید.

اصول پایه ناپیوستگی رگرسیون

جدول ۷.۱، طرح پایه و واریته‌های آن را به طور خلاصه نشان می‌دهد. در این فصل همچنین به بررسی نحوه اتصال به این طرحها آزمایش‌های تصادفی و شبه‌آزمایش‌ها خواهیم پرداخت. در اینجا کار را با توصیف طرح RD پایه شروع می‌کنیم.

ساختار پایه

در طرح RD، محقق باید شرکت‌کنندگان را به دو یا چند شرایط مداخله تخصیص داده و سپس یک پس‌آزمون انجام می‌دهد. محقق، افراد را بر اساس یک نمره برش روی متغیر مبنای تخصیص، به شرایط مختلف تخصیص می‌دهد. در اینجا، این کار- مانند آنچه در آزمایش تصادفی اتفاق می‌افتد- با انداختن سکه یا قرعه انجام نمی‌شود. **متغیر تخصیص** می‌تواند هر مقیاسی باشد که قبل از مداخله تعیین می‌شود، و واحدها بر اساس نمره‌ای که در آن مقیاس می‌گیرند، به شرایط مختلف تخصیص داده می‌شوند؛ واحدهایی (افرادی) که نمره آنها در این مقیاس، در یک طرف مقدار برش است، به یک از شرایط (مداخله یا کنترل)، و آنهایی که نمره آنها در طرف دیگر نمره برش قرار دارد، به شرط دیگر تخصیص داده می‌شوند.

جدول ۷.۱. خلاصه‌ای از طرح‌های ناپیوستگی رگرسیونی

---

۱. *طرح پایه.* شرکت‌کنندگان بر اساس اینکه در بالا یا پایین نمره برش هستند، به گروه مداخله یا کنترل تخصیص داده می‌شوند.

- متغیر تخصیص می‌تواند هر متغیری باشد که قبل از مداخله اندازه‌گیری شده است. این شامل پیش‌آزمون صورت‌گرفته روی متغیرهای خروجی نیز می‌شود.
- متغیر تخصیص، لزوماً با خروجی همبستگی ندارد.
- طرح زمانی در قدرتمندترین حالت است که، نقطه برش در محل میانگین متغیر تخصیص باشد.
- بیش از یک متغیر تخصیص را می‌توان به کار گرفت.

---

۲. *واریته‌ها در طرح پایه*

- به جای مقایسه مداخله و کنترل، دو مداخله مقایسه می‌شوند.
- سه شرایط مختلف با استفاده از تخصیص بر مبنای دو مقدار برش مقایسه می‌شوند.

- از دو برش استفاده می‌شود تا بازه[۳۸۰] برش تشکیل شود؛ آنهایی که درون بازه قرار می‌گیرند به یک شرایط، و آنهایی که بیرون بازه برش قرار می‌گیرند به شرایط دیگر تخصیص داده می‌شوند.

۳. *ترکیب ناپیوستگی رگرسیون با تصادفی‌سازی*

- از دو برش استفاده می‌شود، مقادیر درون بازه برش به صورت تصادفی تخصیص داده می‌شوند، مقادیر بالای بازه (یا پایین) را به مداخله، و مقادیر پایین بازه (یا بالا) را به گروه کنترل تخصیص می‌دهند.
- از یک برش استفاده می‌شود؛ در یک طرف نقطه برش تخصیص تصادفی صورت می‌گیرد، و در طرف دیگر برش همه شرکت‌کنندگان به یک شرایط تخصیص داده می‌شوند.
- از بازه‌های برش چندگانه بهره می‌گیرد، که در برخی از بازه‌ها تصادفی‌سازی، و در برخی دیگر از RD استفاده می‌شود.
- از بازه‌های برش متعدد بهره می‌گیرد، که میزان شرکت‌کنندگان تخصیص داده‌شده به گروه مداخله (در طول بازه‌ها) در حال افزایش است.

۴. ترکیب ناپیوستگی رگرسیون و مؤلفه‌های طرحهای شبه‌آزمایشی

- از یک بازه برش استفاده می‌شود؛ بیرون بازه طرح RD ، و درون بازه بصورت تخصیص خود-انتخابی[۳۸۱] است.
- از یک طرح RD پایه استفاده می‌شود، اما در انتهای مطالعه، مداخله برای همه شرکت‌کنندگان کنترل اعمال می‌شود.
- یک پیش‌آزمون دوگانه افزوده می‌شود تا به شناسایی شکلهای تابعی در غیاب مداخله کمک شود.
- یک گروه کنترل همتایان افزوده می‌شود، تا بتوان شکل تابعی را در یک گروه همتا که مداخله دریافت نکرده است، مدلسازی کرد.

طرح پایه را می‌توان به صورت زیر نشان داد:

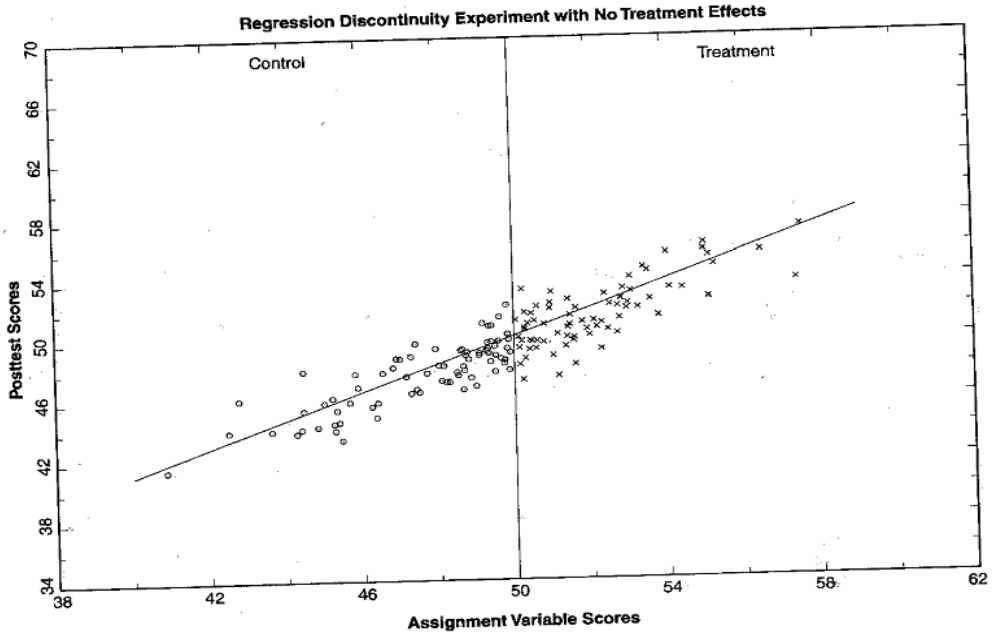
$$O_A \quad C \quad X \quad O_2$$
$$O_A \quad C \quad \quad O_2$$

که در آن، O_A یک اندازه‌گیری پیش از تخصیص، از متغیر تخصیص است، و C نشان می‌دهد که واحدها براساس نمره برش به شرایط، تخصیص پیدا کرده‌اند. یعنی اگر j، نمره برش در هنگام O_A باشد، آنگاه هر نمره بالاتر یا

[۳۸۰] Interval

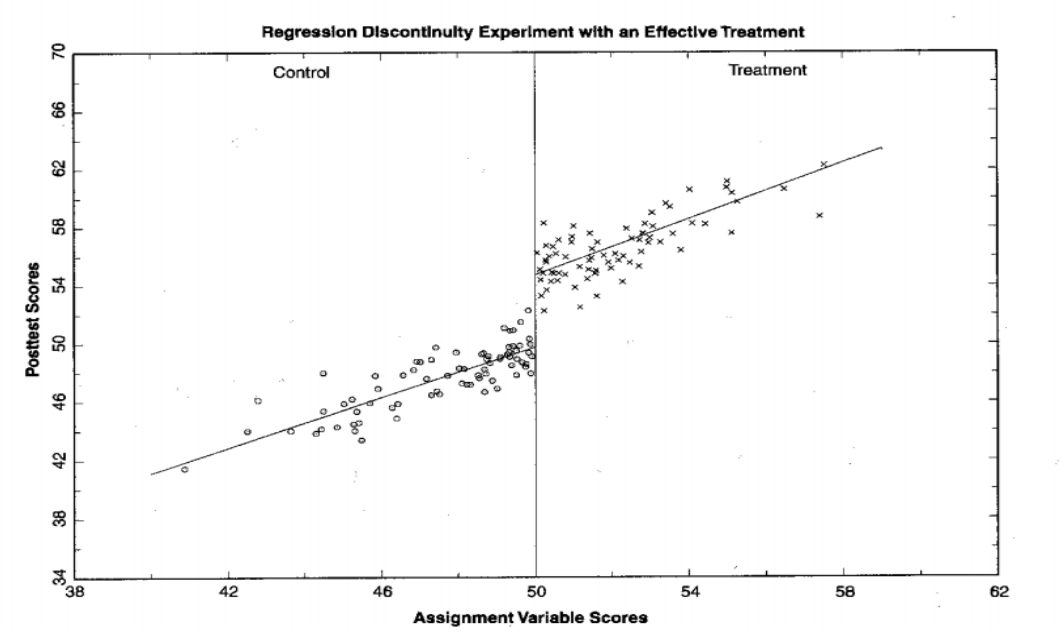
[۳۸۱] Self-selection

مساوی j در یک گروه قرار می‌گیرد، و هر نمره کمتر از j در گروه دیگر قرار می‌گیرد. متغیر تخصیص باید حداقل مشخصات یک مقیاس اندازه‌گیری ترتیبی[۳۸۲] را داشته باشد، یعنی به صورت یکنواخت[۳۸۳] افزایش پیدا کند؛ متغیرهای اسمی واقعی، مانند قومیت، استثناء هستند.



شکل ۷.۱. آزمایش ناپیوستگی رگرسیونی بدون اثرات مداخله

[382] Ordinal
[383] Monotonically



شکل ۷.۲. یک آزمایش ناپیوستگی رگرسیونی با یک مداخله مؤثر

شکل‌های ۷.۱ و ۷.۲، دو نمود مختلف از یک مطالعه RD فرضی را نشان می‌دهند. متغیر تخصیص می‌تواند یک نمره پیش‌آزمون در متغیر وابسته فرضیه باشد، مثلاً یک آزمون پیشرفت تحصیلی یا معیاری از شدت بیماری. شرکت‌کنندگانی که نمره آنها بالاتر از نقطه برش است (در اینجا ۵۰) [۳۸۴] به گروه مداخله، و دیگران به گروه کنترل تخصیص یافته‌اند. هر دو شکل، نمودار پراکندگی نمرات متغیر تخصیص را نسبت به نمرات پس‌آزمون نشان می‌دهند. خط عمود در محل نمره برش، شرکت‌کنندگان گروه مداخله را از گروه کنترل جدا می‌کند. اگر خط تفکیک‌کننده را در نظر نگیریم، این دو شکل همانند هر نمودار پراکندگی دیگریست که نشان‌دهنده یک رابطه مثبت خطی بین دو متغیر باشد. شکل ۷.۱، نتایجی را نشان می‌دهد که در صورت بی‌اثر بودن مداخله، انتظار داریم بدست آوریم، و شکل ۷.۲ نشان می‌دهد که اگر مداخله موثر باشد، نمودار پراکندگی چه تغییری پیدا خواهد کرد. اندازه اثر ترسیم‌شده حدود ۵ واحد است؛ که این مقدار به نمرات پس‌آزمون همه شرکت‌کنندگان گروه مداخله افزوده شده است. خط رگرسیون جابه‌جایی عمودی (یا ناپیوستگی) به میزان حدوداً ۵ واحد در محل نقطه برش در شکل ۷.۲، این مسأله را نشان می‌دهد.

در مطالعات واقعی، متغیر تخصیص معمولاً شایستگی یا نیازها را ارزیابی می‌کند. نمره برش نیز نشان می‌دهد که کسانی که کاری را به خوبی انجام داده‌اند، پاداش دریافت کرده‌اند، و یا کسانی که نیازهای ویژه‌ای داشته‌اند که آنها را واجدشرایط دریافت خدمات کمکی می‌کند، خدمات دریافت کرده‌اند. بر این اساس، RD در مواردی سودمند

[۳۸۴] انتخاب نمره برش به مقیاس متغیر تخصیص بستگی خواهد داشت؛ کاربرد ۵۰ در این داده‌های فرضی به دلخواه صورت گرفته است و بیانگر مقیاس دلخواهی است که در ساخت مثال به کار گرفته شده است.

است که بنا بر گفته منتقدین، تخصیص تصادفی سبب شود افراد شایسته از پاداشی که مستحقش هستند محروم شوند، و یا افراد محتاج به درمان اما ضعیف، از دریافت مداخله باز بمانند (Beecher, 1966، Marquis, 1983؛ Mike, 1989, 1990؛ Schaffner, 1986؛ Veatch & Sollitoo, 1973). اگرچه نقطه ضعف RD در این است که برای اینکه این طرح قدرتی برابر با طرح آزمایش تصادفی داشته باشد، باید شرکت‌کنندگان (اندازه نمونه) بیشتری را بکار بگیرد (Trochim & Cappelleri, 1992؛ Williams, 1990). به عنوان مثال برنامه‌ای آموزشی را فرض کنید که برای جوانان باهوش و بااستعداد طراحی شده، و دانشجویان برای پذیرفته شدن باید نمره بیشتر از ۹۸ بگیرند؛ این مثالی است که در آن از مقیاس شایستگی در طرح RD استفاده شده است. و اگر مقیاس در طرح RD مبتنی بر نیاز بود، باید به دانش‌آموزانی که نمره درس قرائت آنها کمتر از ۲۵ از صد بوده است، آموزش اضافی داده شود. در هر دو مورد، اثر مداخله، یک جابه‌جایی به سمت بالا یا پایین در خط رگرسیونی که تخصیص را به نتایج مرتبط می‌سازد، ایجاد می‌کند (خواه تغییر در میانگینی که در آن، نمره‌های متغیر نتیجه‌ای در یک طرف برش، به میزان میانگین اثر افزایش یافته‌اند، و خواه تغییر در شیبی که در آن، خط رگرسیون در یک طرف برش، از طرف دیگر، شیب بیشتری دارد). این جابه‌جایی در میانگین یا شیب باید دقیقاً در نقطه‌ای رخ دهد که نمره برش متغیر تخصیص در آن تعریف شده است. اساساً نامگذاری این طرحها به «ناپیوستگی رگرسیونی» بواسطه همین جابه‌جایی (یا ناپیوستگی) مربوط به این نقطه در خط رگرسیون است.

مثال‌هایی از طرح‌های ناپیوستگی رگرسیون

اجازه دهید به مثال‌هایی واقعی از کاربرد این طرحها اشاره کنیم. در آموزش جبرانی، به طور معمول همه کودکان تحت پیش‌آزمون -مانند آزمون قرائت- قرار می‌گیرند. آنهایی که نمره‌ای زیر برش دارند، باید دوره قرائت را بگذرانند، و آنهایی که نمره بالاتر دارند نیازی به گذراندن دوره ندارند. سپس همه کودکان در یک آزمون قرائت شرکت می‌کنند (نیازی نیست که مشابه آزمون بکار گرفته شده در پیش‌آزمون باشد)، و تحلیل از نظر ناپیوستگی رگرسیونی انجام می‌شود. تروکیم (Trochim, 1984) داده‌های بدست آمده از نمونه‌های از چنین آموزشهای جبرانی، که متغیر تخصیص آنها نمرات قرائت پیش‌آزمون بودند، را به صورت گسترده تحلیل کرد. به عنوان مثال، تحلیل او از دوره قرائت جبرانی درجه-دوم در پرووِیدنس رودآیلند نشان داد که این دوره، توانایی قرائت دانش‌آموزان را بسیار بهبود داده است. تفسیر جایگزین اصلی در این مورد، شانس است. تروخیم (Trochim, 1984) اذعان می‌کند که بسیاری از دیگر برنامه‌های آموزشی جبرانی که وی مورد بررسی قرار داده، اثرات صفر یا منفی داشته‌اند.

مطالعه مارک و ملور (Mark and Mellor, 1991) به بررسی این موضوع می‌پرداخت که آیا رویدادهایی با ارتباط شخصی بالا، سوگیری پیش‌بینی‌پذیری ماوقع[۳۸۵] (تمایل برای گفتن اینکه نتیجه قابل پیش‌بینی بود، البته بعد از

[385] Hindsight

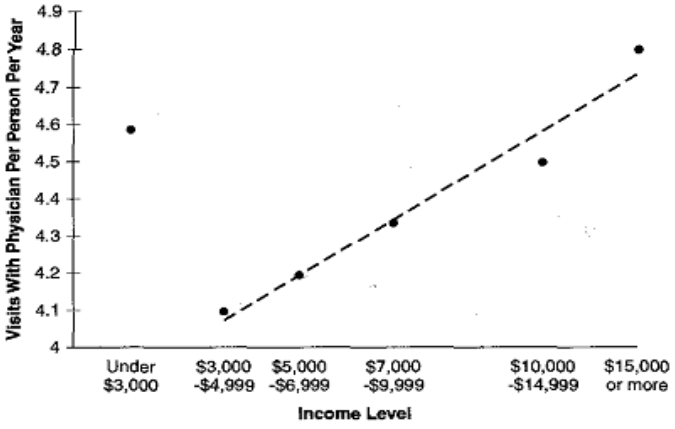
آنکه نتیجه مشخص شد– مثل عبارت من به شما گفته بودم) را افزایش/کاهش می‌دهند؟ طرح RD برای این منظور مورد استفاده قرار گرفت، زیرا در میان کارگرهای اتحادیه شاغل در کارخانه‌های بزرگ مورد مطالعه، آنهایی که ۲۰ سال سابقه کار داشتند، اخراج نشده، اما آنهایی که سابقه کمتری داشتند، اخراج شده بودند. متغیر مستقل از دست دادن شغل، متغیر تخصیص سابقه سرکارگری با نقطه برش ۲۰ سال، و متغیر نتیجه‌ای (وابسته)، درجه‌ی پیش‌بینی‌پذیری[۳۸۶] بود. نتایج (شکل ۷.۳) نشان دادند که کسانی اخراج شده بودند فکر می‌کردند اخراج آنها قابل پیش‌بینی نبوده (من فکر نمی‌کردم که اخراج شوم). توجه داشته باشید که نمودار۷.۳ بر اساس نمرات پاسخهای میانگین‌گروهی[۳۸۷] رسم شده است. این نمرات گروهی از طریق دسته‌بندی‌کردن پاسخ‌دهندگان بر مبنای تعداد سالهای سرکارگری بدست آمده است. علّت این کار آن بود که خروجی آنها یک مقیاس درجه‌بندی ۳-مقداره[۳۸۸] بود، که ناپیوستگی و میانگین گروهی را به شکل بصری نشان نمی‌داد. اگرچه داده‌های آنها در سطح فردی تحلیل شد.

مثالی دیگر، مربوط به مطالعه نحوه تأثیر برنامه مدیکید[۳۸۹] بر ویزیت پزشکان، بعد از ابلاغ برنامه در سال ۱۹۶۴ است (شکل ۷ – ۴؛ Lohr, 1972؛ Wilder, 1972). درآمد خانوارها متغیر تخصیص بود، سطح درآمد واجدشرایط تعیین‌شده در قانون، به عنوان برش انتخاب شد، و فراوانی ویزیت پزشک در سال، متغیر نتیجه‌ای بود. و مانند مثال قبل، به جای نمرات فردی، میانگینهای گروهی برای رسم نمودار بکار گرفته شد.

شکل ۷.۳. اثر اخراج شدن روی سوگیری پیش‌بینی‌پذیری ماوقع

از « Effect of self-relevance of an event on hindsight bias, the foreseeability of a layoff by M. M. Mark and S. Melor, 1991, Journal of Applied Psychology, 76. Pp. 569-577. حق تکثیر American Psychological.

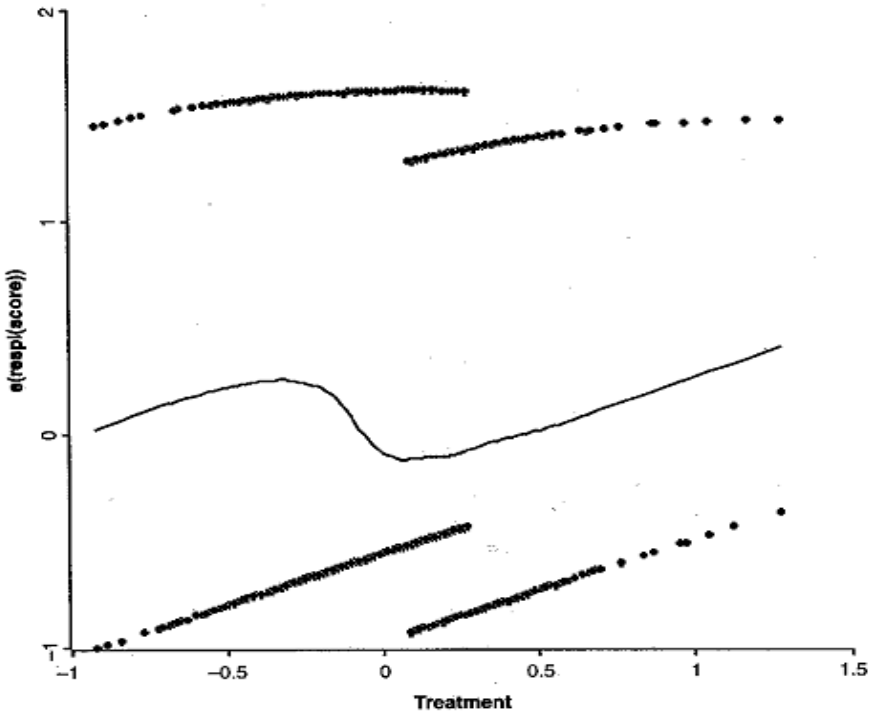
[386] Foreseeability
[387] Group average responses
[388] 3-point rating scale
[389] Medicaid



شکل ۷.۴. تحلیل کمّی شده طرح تنهاپس‌آزمون با گروههای کنترل متعدد برای اثر مدیکید (بر اساس Lohr, 1972؛ Wilder, 1972).

خط رگرسیونی که درآمد را به ویزیت پزشک مرتبط می‌کرد، به طور کلی رابطه‌ای مثبت را نشان می‌داد؛ شاید به این دلیل که افراد با درآمد بالاتر می‌توانند مراقبت پزشکی بیشتری دریافت کنند، و یا اینکه این افراد سن بالاتری دارند، و در نتیجه، از سلامتی کمتری برخوردارند. برای هدف این تحقیق، نتیجه مهم این بود که تعداد ویزیت پزشکان در محل نقطه برشی که واجدشرایط بودن برای برنامه مدیکِید را تعریف می‌کرد، افزایش مشهودی پیدا کرد؛ اثری که به میزان افزایش مشهود در سری‌زمانی راهنمای تلفن سیسیناتی قابل‌توجه بود (شکل ۶.۱ را ببینید). اگرچه اندازه کوچک نمونه، و این حقیقت که تنها یک واحد مداخله مورد مشاهده قرار می‌گیرد، از جمله مسائل این مطالعه به شمار می‌آیند، و هر تفسیر جایگزین احتمالی موجّه باید با در نظرگرفتن فاکتورهایی غیر از مدیکِید، که می‌توانسته‌اند تعداد ویزیتهای پزشکان را برای همان گروه درآمدی (واجد شرایط مدیکِید) افزایش دهند، مورد بررسی قرار گیرد. این کار در ایالات‌متحده قابل‌انجام است، زیرا برنامه‌های اجتماعی متعدد اثرگذار بر تعداد ویزیت پزشکان، درآمد فرد (با نقطه برشی در حدود نقطه برش برنامه مدیکِید) را شرط واجدشرایط بودن برای شمول در برنامه قرار می‌دهند. لور (Lohr, 1972) برای بررسی این تفسیر، داده‌های مربوط به یک سال قبل از اجرای مدیکِید را بررسی کرد، و دریافت که هیچ ناپیوستگی‌ای در محل نقطه واجدشرایطی مدیکِید در آن سال دیده نمی‌شود. با توجه به این نتایج، هیچ برنامه دیگری هم نمی‌توانسته ویزیت دکتر را افزایش دهد، مگر اینکه از نقطه برش مشابهی استفاده کرده باشد، و در همان سال اجرای مدیکِید اجرا شده باشد؛ این احتمالیست که باید موردبررسی قرار گیرد. مثال نهایی، ارزیابی برک و دلیو (Berk and deLeeuw, 1999) از سیستم طبقه‌بندی زندانیان کالیفرنیا است (شکل ۷ – ۵). بر اساس نمره طبقه‌بندی (متغیر تخصیص)، که ترکیبی از چند متغیر مانند طول مدت زندان، سن، و حبس‌های پیشین است، زندانیان به مکان‌هایی با امنیت کم (مداخله) یا امنیت بالا فرستاده شدند، و خروجی، متغیری دومقداره بود عبارت از اینکه آیا زندانی در ۱۸ ماه آینده سوء رفتاری نشان خواهد داد یا خیر. نتایج آماری

نشان داد که قرار گرفتن در مکان‌هایی با امنیت بالا، امکان سوءرفتار زندانیان را تا نصف کاهش می‌دهد. شکل ۷.۵، از چهار جهت با نمودارهای پیشین تفاوت دارد: (۱) خروجی آن دو-مقداره است، بنابراین، نقاط داده‌ها به صورت گروهی در بالای نمودار، و گروه دیگری در پایین نمودار مجزا شده‌اند، و وجود یا عدم وجود سوءرفتار را نشان می‌دهند؛ (۲) نمودار، نقاط داده خام را رسم نمی‌کند، بلکه نمودار متغیرهای-افزوده‌شده[۳۹۰] است (Cook & Weisberg, 1994) که در آن، هم شاخص پاسخ، و هم شاخص مداخله برای متغیر تخصیص باقیمانده‌یابی[۳۹۱] می‌شوند؛ (۳) به جای متغیر تخصیص، یک متغیر مداخله روی محور افقی رسم می‌شود، و (۴) نمودار، شامل یک لُوِس[۳۹۲] است- لُوِس عبارتست از خطی در میانه نمودار، که به منظور خلاصه کردن نقاط داده به صورتی که رابطه میان مداخله و نتایج را مشخص‌تر نماید، کشیده می‌شود (Cook & Weisberg, 1994). ناپیوستگی حول نمره صفر متغیر مداخله در خط لُوِس، نشان‌دهنده اثر مداخله است. نمودارهای ۷.۳ تا ۷.۵ به روشنی نشان می‌دهند که نمودارهای طراحی RD ضرورتاً شکل خاصی ندارند.



شکل ۷.۵. نمودار متغیرهای افزوده‌شده برای مشاهده اثر قرار گرفتن زندانیان در مکان‌هایی با امنیت بالا و پایین بر روی سوء رفتارشان. محور افقی، انتظار دریافت مداخله که منوط به نمره تخصیص است را نشان می‌دهد، و محور عمودی، مقدار انتظار خروجی مشروط به نمره تخصیص را نشان می‌دهد.

---

[390] Added variables

[391] Residualized

[توضیح مترجم: هر گاه میان دو متغییر هم‌خطّی (collinearity) وجود داشته باشد، برای برطرف کردن آن، باقیمانده‌یابی انجام می‌شود. به این معنی که میان متغیر اول و دوم رگرسیون گرفته می شود، و باقمیانده این معادله رگرسیون، به جای متغیر اول در معادله اولیه استفاده می شود.]

[392] Lowess Smother

الزامات ساختاری طرح

متغیر تخصیص و برش

در طرح RD، تخصیص به مداخله باید تنها بر اساس نمره برش باشد. از این نظر، متغیر تخصیص در طرحهای RD، به اندازه تخصیص تصادفی در آزمایشهای تصادفی، قطعی و غیرقابل‌اغماض است. نمی‌توان با حذف شرکت-کنندگان از گروهی که وی بر اساس نمره برش (یا پرتاب سکه) به آن تخصیص داده شده، مکانیسم تخصیص را نقض کرد. با توجه به اینکه متغیر تخصیص، و نمره برش، نقشی حیاتی ایفا می‌کنند، چگونه باید آنها را انتخاب کرد؟

متغیر تخصیص نمی‌تواند منتج از مداخله باشد. این شرط در مورد هر متغیر تخصیصی که قبل از شروع مداخله اندازه‌گیری شود، و یا هر متغیری که هرگز تغییر نمی‌کند (مانند سال تولد فرد) برقرار است (Judd & Kenny, 1981a). متغیر تخصیص می‌تواند پیش‌آزمونی که روی متغیر وابسته انجام شده، باشد (در مثال آموزش جبرانی، هم پیش‌آزمون و هم پس‌آزمون، آزمون‌های قرائت هستند). با این وجود، متغیر تخصیص نباید ضرورتاً یک پیش-آزمون باشد. به عنوان مثال، در مطالعه کنترل جرم که خروجی دو-مقداره آن، تکرار یا عدم تکرار جرم بود، تعداد ساعات کار در زندان به عنوان متغیر تخصیص انتخاب شده بود. متغیر تخصیص ممکن است کاملاً بی‌ربط به نتایج بوده، و هیچ معنای خاصی نداشته باشد. به عنوان مثال، کین (Cain, 1975) پیشنهاد می‌کند می‌توان از ترتیب زمانی پرکردن فرم تقاضای شرکت در یک دوره توسط شرکت‌کنندگان، به عنوان معیاری برای تخصیص استفاده کرد؛ به این صورت که، ۲۰ متقاضی‌ای که اول درخواست می‌دهند، در گروه مداخله قرار گیرند، و ۲۰ نفر بعدی در گروه کنترل. دلوس (Deluse, 1999) این کار را در یک مطالعه با طرح RD انجام داد، و زوج‌ها را بر اساس تاریخی که درخواست طلاق داده بودند، به برنامه اجباری آموزش‌های مرتبط با طلاق تخصیص داد.[۳۹۳] در این مثال، RD

[۳۹۳] این مطالعه و در کل کاربرد مرتبه ورود به عنوان متغیر تخصیص، کاملاً مشابه طرح سری زمانی مختل شده است که در آن، مداخله در نقطه خاصی صورت می‌گیرد. با این وجود، این دو طرح در این شرایط متفاوت هستند (اما همیشه اینگونه نیست). معمولاً شرکت‌کنندگان در RD، مستقل از همدیگر هستند؛ اما شرکت‌کنندگان ITS در طول زمان، یکسان هستند (یا حداقل یک زیرمجموعه از آنها در طول زمان یکسان است) بنابراین، نقاط در طول زمان خودهمبسته هستند که این اتفاق در RD رخ نمی‌دهد. اما شرکت‌کنندگان در ITS می‌توانند کاملاً مستقل باشند دقیقاً همانند RD. باز هم در گرافی که مربوط به طرح سری زمانی مختل شده است (مثلاً شکل ۶ – ۱)، نقاط داده که قبل از مداخله ایجاد می‌شوند، پیش‌آزمون هستند یعنی از نظر فیزیکی قبل از شروع مداخله رخ داده‌اند. در طرح ناپیوستگی رگرسیونی که از ترتیب ورود به عنوان متغیر تخصیص بهره می‌گیرد، نقاط داده قبل از برش (مثلاً شکل ۷ – ۲) پس‌آزمون‌هایی هستند که معمولاً بعد از شروع مداخله بدست می‌آیند. با این وجود، می‌توان یک طرح RD را در نظر گرفت که در آن، تمایز این پس‌آزمون‌ها، اهمیت عملی نداشته باشد، اگر مداخله (و کنترل) چنان مختصر باشند که پس‌آزمون به سرعت بعد از تخصیص به شرایط ایجاد شود به گونه‌ای که همه پس‌آزمون‌های کنترل

مانند یک آزمایش تصادفی عمل می‌کند، و فرایند تخصیص (یا پرتاب سکه) به طور میانگین با نتایج ارتباطی ندارد. گرچه، طرحهای RD با وجود یا عدم وجود رابطه میان متغیر تخصیص و متغیر خروجی، همچنان عمل می‌کنند.
بهترین متغیر تخصیص، متغیریست که پیوسته[۳۹۴] باشد، مانند فشار خون در مطالعات پزشکی، درآمد سالانه در مطالعات آموزش شغلی، یا نمره امتحان در تحقیقات آموزشی. این متغیرها، شانس مدلسازی صحیح خط رگرسیون برای هر گروه را بیشنیه می‌کنند؛ چیزی که نقش حیاتی در موفقیت طرحهای RD دارد. بالعکس، نمی‌توان از یک متغیر دو-مقداره مانند جنسیت یا وضعیت سیگاری یا غیرسیگاری بودن، برای تخصیص استفاده کرد. در متغیرهای دوبخشی، تنها یک نمره تخصیص در زیر برش (غیرسیگاری) و یک نمره در بالا (سیگاری) وجود دارد، و هیچ خط رگرسیونی برای هیچکدام از شرایط بدست نمی‌آید؛ و همبستگی بالای متغیر تخصیص با متغیر مجازی[۳۹۵] مداخله، منجر به وابستگی خطی متغیرهای پیش‌بین‌ها می‌شود.

انتخاب نقطه برش

در انتخاب نقطه برش باید ملاحظات فراوانی را در نظر گرفت. این نقطه را می‌توان بر اساس دانش پایه‌ای موجود انتخاب کرد، مثلاً نظر حرفه‌ای در مورد اینکه چه کسی به درمان پزشکی نیاز دارد، یا کدام کودک نیازمند آموزش جبرانی است. اگر برش، میانگین نمرات متغیر تخصیص باشد، هم توان آماری و هم برآورد برهم‌کنش‌ها[۳۹۶] تسهیل می‌شود. با این وجود، اگر شرکت‌کنندگان به کندی، و در طول زمان در مطالعه وارد شوند، به گونه‌ای که نتوان تا زمانی که همه واحدها موجود شوند، میانگینی تعیین کرد، یا اگر هزینه یا شایستگی یا نیاز، اجرای مداخله را به کسانی که نمرات خیلی بالا یا پایینی در متغیر تخصیص بدست آورده‌اند، محدود سازد، تعیین برش برابر با میانگین، امکان‌پذیر نیست. در این حالت، جایابی نقطه برش در محل یک مقدار حدی[۳۹۷]، بر توانایی مدلسازی خط رگرسیون- حتی در صورت وجود متغیر تخصیص پیوسته- آسیب وارد می‌کند. اگر گرنت آموزشی تنها به افرادی داده شود که نمره آنها بالای ۹۹/۹ درصد در یک آزمون است، آنگاه نقاط خیلی کمی در بالای نقطه برش وجود خواهند داشت، و نمی‌توان رگرسیون را مدلسازی کرد، و قدرت آماری نیز به شدت کاهش می‌یابد. وقتی که متغیر تخصیص، چندمقداره باشد -مانند مقیاس لیکرت هفت‌نقطه‌ای- نحوه جایابی نقطه برش اهمیت اساسی پیدا می‌کند. وجود تنها ۱ یا ۲ نقطه (مثلاً ۶ و ۷) در یک طرف برش، برآورد خط رگرسیون را با مشکل مواجه می‌کند.

قبل از هر گونه مداخله روی شرکت‌کنندگان گرفته شوند. اگر اینگونه باشد طرحی که همه این شرایط را براورده کند و نقاط داده کاملاً مستقل داشته باشد، هم RD و هم ITS است.

394 Continuous
395 Dummy
396 Interactions
397 Extreme

حادترین مثال از این حالت در شکل ۷.۴ نشان داده شده است؛ که در آن، تنها یک واحد در شرایط مداخله مورد اندازه‌گیری قرار گرفته، و نمی‌توان هیچ خط رگرسیونی را برای سمت چپ نقطه برش مدلسازی کرد.[۳۹۸]

می‌توان به جای یک متغیر، همزمان از متغیرهای تخصیص متعددی استفاده کرد. به عنوان مثال، اگر مقیاس لیکرت هفت نقطه‌ای به کار رود اما برش بیش از اندازه پایین باشد، می‌توان تکرارهای[۳۹۹] متعدد از متغیر را میانگین‌گیری کرد، تا یک مقیاس با درجه‌بندی جزئی‌تر ایجاد شود. به عنوان مثال، می‌توان از درجه‌بندی‌های چهار پزشک که هر کدام از مقیاس لیکرت هفت‌نقطه‌ای برای تعیین زمان جراحی استفاده کرده‌اند، میانگین گرفت، تا یک متغیر تخصیص با تعداد بازه‌های بسیار بیشتر بدست آید. یا اگر متغیرهای تخصیص مختلف، مقیاسهای متفاوتی داشته باشند، می‌توان ابتدا آنها را استاندارد کرده، و به هر یک وزن متفاوتی اختصاص داد، و سپس از مجموع آنها نمره کل بدست آورد (Judd & Kenny, 1981a؛ Trochim, 1984, 1990). آنگاه شرکت‌کنندگان با استفاده از نقطه برش در نمرات کل، به گروه‌ها تخصیص داده می‌شوند. این میانگین‌ها و نمرات کل می‌توانند در صورتی که متغیرهای تخصیص موجود، خصوصیات توزیعی ضعیفی داشته، و مدلسازی رگرسیونی دقیق ناممکن باشد، به اصلاح مشکلات کمک کنند. استفاده از این استراتژی‌های تخصیص پیچیده، با کاهش همبستگی میان متغیر تخصیص، و دریافت (عدم‌دریافت) مداخله، قدرت طرح را افزایش می‌دهد (Cappelleri & Trochim, 1995؛ Judd & Kenny, 1981a). یا اینکه به جای ترکیب شاخص‌های شایستگی یا نیاز مجزا، می‌توان روی هر یک از این متغیرها نقطه برشی تعیین کرد، و شرکت‌کنندگانی که در تعداد معینی از این شاخص‌ها (مثلاً حداقل در ۶ شاخص از ۱۲ شاخص، بالای برش باشند)، و یا در همه متغیرهای تخصیص به صورت همزمان بالای برش باشند، در مداخله شرکت داده شوند. وقتی این استراتژی‌های پیچیده مورداستفاده قرار می‌گیرند، می‌توان از قوائد تحلیلی خاصی بهره گرفت (Judd & Kenny, 1981a؛ Trochim, 1984, 1990 را ببینید).

تخصیص به مداخله باید کنترل شده باشد؛ اینکار کاربردهای پس‌نگر طرح را بی‌اثر می‌نماید. مثالی از این حالت، طرحی شبه‌آزمایشی با گروه مقایسه غیرهم‌ارز کنترل‌نشده است، که در آن، محقق همه شرکت‌کنندگان گروه مداخله را که از نظر نمره کسب‌شده در متغیر برش، در طرف گروه کنترل قرار گرفته‌اند، و همینطور همه شرکت-کنندگان کنترلی که از نظر نمره کسب شده در متغیر تخصیص، در سمت مداخله قرار گرفته‌اند، را حذف می‌کند (Judd & Kenny, 1981a). در اینجا، مکانیسم تخصیص اولیه شناخته‌شده نیست، کنترل کمتری روی آن وجود دارد، که این می‌تواند منجر به بروز سوگیری انتخابی گردد، که قابل‌حل با طرحهای RD نیست. اینگونه حذف

---

[۳۹۸] اگر بتوان خط رگرسیون را برای گروه مقایسه به دقت مدلسازی کرد، آنگاه می‌توان بررسی کرد که آیا میانگین گروه مداخله، با تصویر خط گروه مقایسه اختلاف زیادی دارد یا خیر (از طریق تفریق میانگین گروه مداخله (به جای نمرات برش) از نمرات پیش‌آزمون).

[399] Replication

مواردیکه به طور نادرست تخصیص داده‌اند، سبب ایجاد انحنا[۴۰۰] در تابع رگرسیون صحیح می‌شود، که می‌تواند با اثر مداخله اشتباه گرفته شود (Goldberger, 1972a).

دیگر الزامات

مهم است بتوانیم شکل تابع کلی که متغیرهای تخصیص و خروجی را به یکدیگر مرتبط می‌کند، را شناسایی (یعنی اینکه آیا خطی، منحنی، چرخه‌ای و ... است). یک مدل چندجمله‌ای[۴۰۱] می‌تواند برای توصیف این شکل مناسب باشد[۴۰۲] (Trochin, 1984)، همچنین برخی دیگر تبدیلات را می‌توان برای تخصیص یا متغیر پس‌آزمون به کار گرفت (مانند تبدیل لگاریتمی). این توابع می‌توانند به‌خوبی اکثر روابط را توصیف کنند. اما اگر شکل تابعی در تحلیل درست انتخاب نشود، آثار مداخله همراه با سوگیری برآورد می‌شوند.

همه شرکت‌کنندگان باید پیش از آنکه به یک گروه تخصیص داده شوند، عضو جمعیت موردنظر باشند، اگرچه نحوه تعریف جمعیت در طرحهای RD روشن نیست. تعریفی مانند آنچه در مدل علّی رابین درمورد آزمایش‌های تصادفی به کار رفته است (Rubin, 1974, 1977, 1978, 1986؛ Holland, 1986)، را می‌توان برای RD هم به کار گرفت. بر اساس مدل رابین، قبل از تخصیص تصادفی، باید همه واحدهای یک آزمایش قابلیت دریافت مداخله را داشته باشند. بنابراین، در RD نیز، اگر برش به گونه دیگری تنظیم شود، همه واحدهای مطالعه باید قادر به دریافت مداخله باشند. به عنوان مثال، فرض کنید که در مدرسه الف برای دانش‌آموزانی که نمرشان بالاتر از برش است، یک مداخله خاص اجرا می‌شود، و گروه کنترل شامل دانش‌آموزانی از مدرسه ب است که نمراتی پایین‌تر از نقطه برش دارند. از آنجا که این مداخله در مدرسه ب انجام نمی‌شود، دانش‌آموزان مدرسه ب حتی اگر نمره‌ای بالاتر از برش داشته باشند نیز نمی‌توانند مداخله را دریافت کنند. به علاوه، بودن (انتخاب شدن برای حضور) در مدارس الف و ب با متغیرهایی غیر از متغیر برش صورت می‌گیرد که برای محقق نامعلوم هستند. این باعث بروز نوعی سوگیری انتخاب غیرقابل‌کنترل در طرحهای RD می‌شود.

در حالت ایده‌آل، همانند آزمایش تصادفی، همه آنهایی که در مداخله هستند باید یک به میزان برابری مداخله دریافت کرده باشند، و افراد در گروه کنترل هیچ مقداری از مداخله را دریافت نکنند. با این وجود، برخی شرکت-کنندگان گروه مداخله، ممکن است مقدار کمتری از مداخله را (به نسبت به دیگران) دریافت کنند، یا ممکن است انتشار مداخله سبب شود که برخی از افراد گروه کنترل به گروه مداخله شبیه شوند. حذف این شرکت‌کنندگان یکپارچگی تخصیص را تضعیف می‌کند. بنابراین، هم در RD، و هم در آزمایشهای تصادفی، معمولاً همه شرکت-

---

[400] Curvilinearity

[401] Polynomial

[۴۰۲] به عنوان مثال، معادله‌ای که شامل $X, X^2, X^3, ..., X^n$ باشد.

کنندگان در شرایطی که به آن تخصیص داده شده‌اند، باقی می‌مانند (Pocock, 1983 ؛Lavori, 1992)[۴۰۳]. همچنین فرض کنید که درصد اجرای مداخله به صورت سیستماتیک با نمره متغیر تخصیص کوواریانس داشته باشد، به گونه‌ای که کسانی که نمرات پایین‌تری گرفته‌اند، مداخله کمتری دریافت کنند. این منجر به تولید یک اثر برهم-کنشی مصنوعی می‌شود. به این معنی که مداخله‌ای که برای همه شرکت‌کنندگان به یک اندازه مؤثر است، ممکن است به نظر برسد روی افرادی که نمره پیش‌آزمون کمتری دارند، تأثیرگذارتر است.

تغییر در طرح پایه

اکنون که اصول پایه مربوط به RD مشخص شده‌اند، به سادگی می‌توان این ایده‌ها را به واریته‌های پیچیده‌تر بسط داد. برخی از این تغییرات عبارتند از:

- مقایسه دو مداخله با همدیگر به جای مقایسه مداخله با کنترل؛ به غیر از این مورد، طراحی و هم تحلیل همانند مشابه قبل بوده و تغییری نمی‌کند؛
- می‌توان سه شرایط، مثلاً مداخله استاندارد، نوآوری و کنترل را بررسی کرد، و شرکت‌کنندگان بر اساس برش‌های متعدد تخصیص یابند (دو برش برای سه شرایط، سه برش برای چهار شرایط، و الی آخر).
- مداخلهها را می‌توان برای گروه‌های مختلف با دوزهای متفاوت اجرا کرد، به گونه‌ای که کسانی که نیاز بالاتری دارند، دوز بالاتری دریافت کنند[۴۰۴].
- حتی اگر دو شرایط اجرا شده باشد، می‌توان دو نمره برش روی متغیر تخصیص در نظر گرفت تا شرکت‌کنندگان به سه گروه تقسیم‌بندی شوند. گروه میانی به یکی از شرایط تخصیص می‌یابد- مثلاً مداخله- و دو گروه دیگر، شرایط دیگری دریافت می‌کنند. اگر مداخله مؤثر باشد، و کنترل تأثیری نداشته باشد، رابطه بین متغیر تخصیص و متغیر دو-مقداره اندازه‌گیری‌کننده در شرایط مداخله، به شکل منحنی خواهد بود. این حالت بوسیله کاهش هم‌خطّی میان این دو متغیر خروجی در مدل خطی، قدرت طرح را افزایش می‌دهد.

---

[۴۰۳] تحلیل‌هایی که به تازگی برای بررسی این مشکل در آزمایش‌های تصادفی توسعه یافته‌اند (Rubin, 1992a ؛Angrist, Imbens & Rubin, 1996a) را می‌توان برای طرح RD نیز به کار برد.

[۴۰۴] در این تغییرات ثانویه و ثالث، تحلیل استانداردی که برای طرح پایه توصیف می‌کنیم باید شامل متغیرهای فرضی چندگانه Z هم باشد تا بتواند هر مداخله را در معرض قیدهای کدگذاری متغیر فرضی نشان دهد. با این وجود، تنها یک برش را می‌توان از متغیر تخصیص مدل تفریق کرد (Trochim, 1984؛ صفحات ۱۳۴ – ۱۳۵).

نویسندگان مختلف (مانند Judd & Kenny, 1981a؛ Trochim, 1984) توصیه‌هایی در مورد اجرا و تحلیل این تغییرات، و دیگر انواع تغییرات در طرحهای RD پایه، ارائه کرده‌اند.

کاربردهای متنوعی برای استراتژی‌های تخصیص مبتنی بر برش وجود دارد (Atwood & Taylor, 1991). در اینجا چهار مثال ارائه می‌شود. اول اینکه رابین (Rubin, 1977) پیشنهاد می‌کند در انتهای آزمایش تصادفی کسانی که نمره‌شان در متغیر خروجی زیر نقطه برش قرار می‌گیرد، جلسات تقویتی دریافت کنند، یا کسانی که نمره آنها کمتر از مقدار برش مربوط به یک مداخله است، می‌توانند توجه بیشتری بدست آورند. مثال دوم، اصلاحات بهبود آزمایشگاه‌های بالینی است، که توسط کنگره ایالات متحده در سال ۱۹۸۸ تصویب شد، تا نظارت بر آزمایشگاه‌های پزشکی اجباری شود. آزمایشگاه‌هایی که در زیر نمره ارزیابی شایستگی قرار می‌گیرند، باید اصلاحات خاصی را از مراکز کنترل بیماری (CDC) فرا گرفته، و اجرا کنند. ارزیابی این برنامه را می‌توان با استفاده از طرحهای RD انجام داد. مثال سوم، در یک مطالعه تحقیقاتی با استفاده از ۲ میلی‌گرم دوزی از صمغ نیکوتین برای جلوگیری از بازگشت به سیگار رخ داد. سازنده دارو، از دوز ۴ میلی‌گرم استفاده کرده بود، و اعتقاد داشت که این دوز برای اغلب سیگاری‌ها، بیش از اندازه قوی است. راه حل این کار، تخصیص دوز ۴ میلی‌گرمی به سیگاری‌هایی با نمره بالاتر از یک مقدار معین در مقیاس اندازه‌گیری اعتیاد به سیگار (مثلاً Fagerstrom, 1978)، و دوز ۲ میلی‌گرمی به افراد زیر برش بود. به عنوان مثال چهارم، برنامه پیش‌دبستانی، کودکان را بر اساس نیازشان به خدمات پیش‌دبستانی، درجه‌بندی می‌کند. این درجه‌بندی در تصمیم‌گیری برای اینکه چه کسی خدمات دریافت کند کاربرد خواهد داشت. اگر این درجه‌بندی تنها فاکتور اثرگذار بر تقسیم‌بندی باشد، داشتن یک طرح RD امکانپذیر خواهد بود.

RD می‌تواند نسبت به گذشته کاربرد بسیار بیشتری داشته باشد. آزمایش‌های تصادفی نیز در سال ۱۹۲۰ معرفی شدند، اما تا سال ۱۹۵۰، توفیق چندانی در جلب نظر متخصصان علوم اجتماعی و بهداشتی نداشتند. شاید این تأخیر ۳۰ ساله بین معرفی و کاربرد، برای طرحهای RD نیز مصداق داشته باشد، و بتوان گفت که این طرحها در ۳۰ سال آینده با فراوانی بیشتری بکار گرفته خواهند شد.

نظریه طرح ناپیوستگی رگرسیون

بسیاری از خوانندگان فکر می‌کنند ممکن نیست که طرح ناپیوستگی رگرسیون برآوردهای مفید و فاقد سوگیری (یا با سوگیری خیلی اندک) از اثرات مداخله ارائه کند. در این بخش، شرح خواهیم داد که چرا این طرح می‌تواند موفق‌آمیز باشد. در بخش اول توضیحات نشان می‌دهیم که آزمایش‌های تصادفی نیز از ناپیوستگی‌های رگرسیونی برای برآورد اثرات بهره می‌گیرند؛ و قسمت دوم، ناپیوستگی رگرسیونی را به شرایطی ارتباط می‌دهد که تحت آن می‌توان سوگیری انتخاب را در هر شبه‌آزمایشی به صورت موفقیت‌آمیز مدلسازی کرد (در مدل لحاظ کرد).

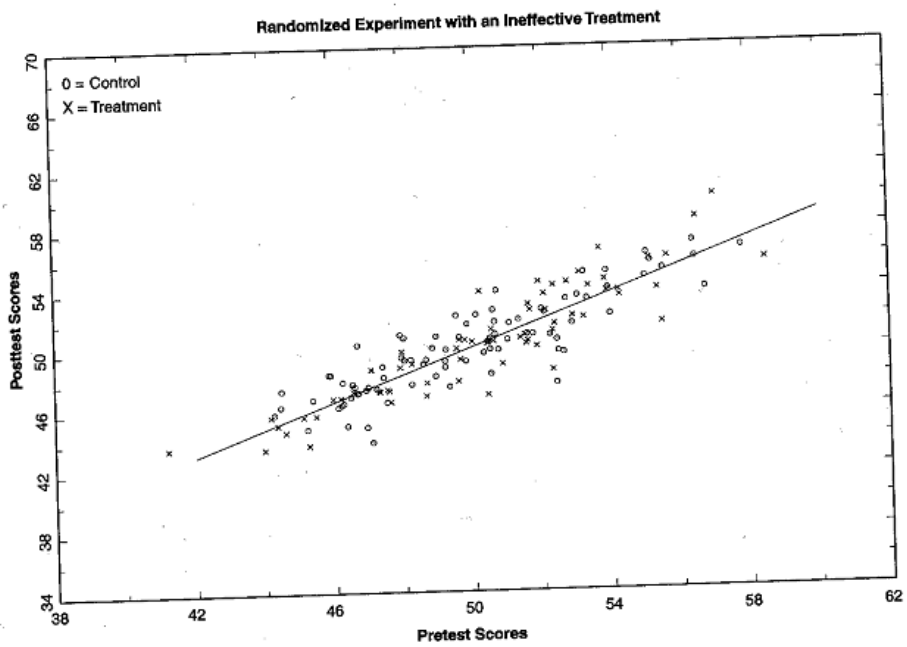
ناپیوستگی‌های رگرسیونی به عنوان اثرات مداخله در آزمایش تصادفی

فرض کنید که شرکت‌کنندگان با همان روندی که برای شکل‌های ۷.۱ و ۷.۲ مشاهده شد، تحت پیش‌آزمون قرار گرفته‌اند، به صورت تصادفی به گروه مداخله یا کنترل تخصیص یافته‌اند، و سپس تحت پس‌آزمون قرار گرفته‌اند. اگر مداخله مؤثر نباشد، آزمایش تصادفی، همانند شکل ۷.۶، یک نمودار پراکندگی از نمرات پیش‌آزمون بر حسب نمرات پس‌آزمون تولید می‌کند، که خط رگرسیون آن، همانند شکل‌های ۷.۱ و ۷.۲ مثبت است. اما شکل ۷.۶ از دو نظر با شکل ۷.۱ تفاوت دارد. اول اینکه خط برشی ندارد، زیرا تخصیص در طول همه نمرات پیش‌آزمون- به جای اینکه بر اساس نمره برش باشد- تصادفی است. دوم اینکه در شکل ۷.۱، همه شرکت‌کنندگان در مداخله، در سمت راست نمره برش قرار می‌گیرند، و همه شرکت‌کنندگان گروه کنترل در سمت چپ قرار دارند، اما در شکل ۷.۶، شرکت‌کنندگان مداخله و کنترل به صورت تصادفی مخلوط شده‌اند، زیرا تخصیص تصادفی تضمین می‌کند که هیچ رابطه سیستماتیکی بین عضویت در گروه مداخله، و نمرات پیش‌آزمون وجود ندارد.

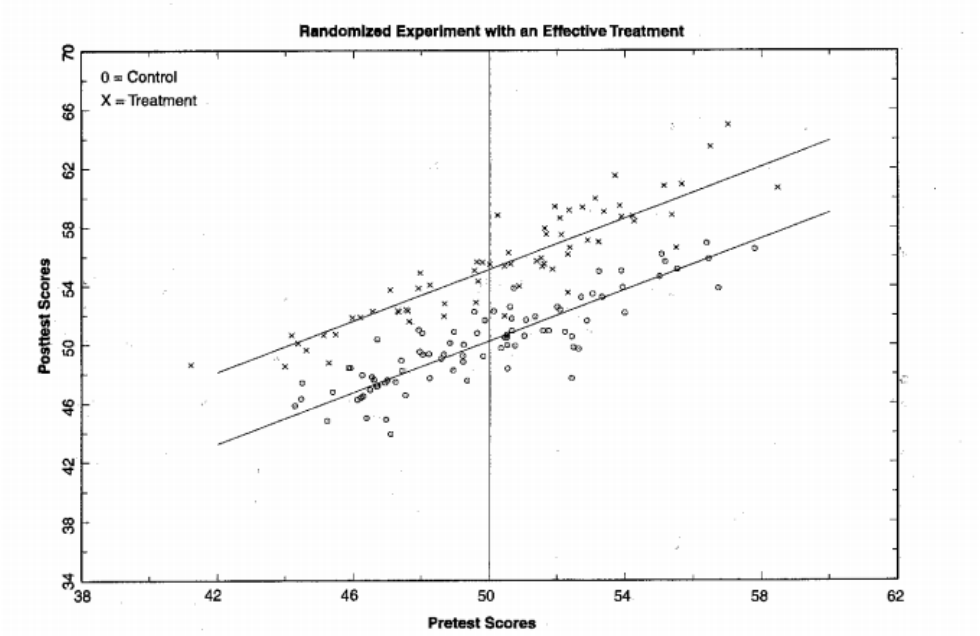
اگر مداخله در آزمایش تصادفی دارای مقیاسی ۵تایی، مشابه با آنچه در طرحهای RD استفاده شد باشد، نمودار پراکندگی‌ای شبیه شکل ۷.۷ بدست می‌آید. اکنون می‌توانیم دو خط رگرسیون رسم کنیم (به ازای هر یک از شرایط). هر دو خط، شیب مثبتی را میان نمرات پیش‌آزمون و پس‌آزمون نشان می‌دهند؛ اگرچه، خط رگرسیون گروه مداخله، ۵ نقطه بالاتر از خط گروه کنترل قرار دارد. توجه داشته باشید که این نمودار تا چه اندازه به نمودار ۷.۲ از طرح RD شبیه است. در واقع، اگر به شکل ۷.۷ در نقطه نمره ۵۰ در پیش‌آزمون، یک خط عمودی اضافه کنیم، جابه‌جایی خط رگرسیون در نمره برش، برآوردی بدون سوگیری از اثر خواهد بود[۴۰۵]. همچنین اگر در شکل ۷.۷، همه شرکت‌کنندگان گروه مداخله در سمت چپ خط برش، و همه شرکت‌کنندگان گروه کنترل در سمت راست را حذف کنیم، گراف حاصله، مشابه نمودار ۷.۲ خواهد شد.

این حذف، اختلاف مهم دیگری بین طرح RD و آزمایش تصادفی را نشان می‌دهد. در شکل‌های ۷.۶ و ۷.۷، میانگین‌های پیش‌آزمون گروه‌های کنترل و مداخله، تقریباً برابر، و در حدود ۵۰ است. زیرا این گروه‌ها با استفاده از تخصیص تصادفی، و به صورت احتمال-محور همسان شده‌اند. در مقابل، در تخصیص مبتنی بر برش، گروه‌هایی تولید می‌شوند که بیشترین میزان تفاوت در میانگین پیش‌آزمونشان وجود دارد (توزیع پیش‌آزمون دو گروه بدون‌هم‌پوشانی است). چگونه ممکن است طرح RD بتواند برآوردی بدون سوگیری از اثرات مداخله حاصل کند، وقتی مقادیر میانگین پیش‌آزمون‌هایش از ابتدا تا این حد متفاوت هستند! با توجه به این اختلافات قابل‌توجه در هم‌ارزی گروهها در پیش‌آزمون، در طرح RD و آزمایش تصادفی، در ادامه شرح خواهیم داد که چرا این نقد، اگرچه منطقی به نظر می‌رسد، اما غلط است؛ و چرا این طرح می‌تواند برای انجام استنباط علّی از هر طرح دیگری-جز آزمایش تصادفی- قوی‌تر باشد.

[۴۰۵] در واقع، میانگین وزنی این ناپیوستگی در کل محدوده مقادیر پیش‌آزمون، یک براوردگر سنتی از اثر مداخله در آزمایش تصادفی است.



شکل ۷ – ۶. آزمایش تصادفی با یک مداخله غیرمؤثر



شکل ۷ – ۷. آزمایش تصادفی با یک مداخله موثر

در آزمایش‌های تصادفی، اثر مداخله از طریق مقایسه میانگین پس‌آزمون گروه مداخله و میانگین پس‌آزمون گروه کنترل برآورد می‌شود. پیشفرض اصلی در اینجا این است که گروه‌های در حال قیاس بواسطه بکارگیری تخصیص

تصادفی، به طور احتمال-محور[406] هم‌ارز هستند (اگرچه این حالت در پس‌آزمون لزوماً برقرار نیست). با این وجود، در طرحهای ناپیوستگی رگرسیونی، مقایسه بین میانگین‌ها صورت نمی‌گیرد، بلکه خطوط رگرسیون با هم مقایسه می‌شوند (مقایسه خط رگرسیون گروه مداخله با خط رگرسیون گروه کنترل که با استفاده از نمرات دو طرف نقطه برش بدست می‌آیند). اگر اثر مداخله وجود نداشته باشند، به جای اینکه فرض کنیم میانگین‌های پیش‌آزمون هم‌ارز هستند، فرض می‌کنیم که شکل تابعی در دو طرف برش هم‌ارز است (خطوط رگرسیون عرض، شیب و صفات یکسانی دارند).

ناپیوستگی رگرسیون به مثابه مدلی کامل از فرایند انتخاب

در اغلب شبه‌آزمایش‌های دیگر که در آنها، تخصیص به مداخله کنترل نشده است، فرایند انتخاب کاملاً ناشناخته است. گاهی شناختی جزئی وجود دارد، اما این فرایند تقریباً هرگز به صورت کامل شناخته شده نیست. به عنوان مثال، محقق چگونه می‌تواند نقش انگیزه، توانایی، وضعیت اجتماعی-اقتصادی و دیگر متغیرها را در تعیین اینکه چه کسی وارد گروه مداخله شود، شناسایی کند؟ و حتی زمانی که محقق بخشی از این متغیرهای انتخاب را می‌شناسد، به ندرت می‌تواند آنها را به طور کامل اندازه‌گیری کند. اگر فرایند انتخاب کاملاً قابل‌درک و اندازه‌گیری باشد، آنگاه می‌توان تفاوتهای انتخاب را به گونه‌ای تعدیل کرد که برآوردی بدون سوگیری از اثر مداخله بدست دهد. در حالت نظری، این شرایط هم در RD، و هم در آزمایش تصادفی فراهم است، و بنابراین هر دو طرح را می‌توان به صورت مواردی خاص (موفق) از مدلسازی سوگیری انتخاب در نظر گرفت (نگاه کنید به پیوست ۵.۱). در یک آزمایش تصادفی، مکانیسم تخصیص کاملاً مشخص است، و عبارتست از پرتاب سکه. در RD نیز این مکانیسم کاملاً مشخص است، زیرا بسته به اینکه نمره در متغیر تخصیص بالا یا پایین نقطه برش باشد، تعیین می‌شود. در هیچکدام از این دو طرح، متغیری نامشخص بر تخصیص شرکت‌کننده به یکی از گروهها (کنترل یا مداخله) اثر نمی‌گذارد. در هر دو مورد، مکانیسم تخصیص را می‌توان به طورکامل اندازه‌گیری و اجرا کرد (یعنی محقق می‌تواند به درستی ثبت کند که آیا نتیجه پرتاب سکه، پشت بوده است یا رو، یا اینکه امتیاز شخص بالای برش بوده است یا پایین آن). البته، احتمال بروز خطاهایی در ثبت، و اثرگذاری فرایندهای اجتماعی برتخصیص در آزمایش‌های تصادفی و RD وجود دارد (Dennis, 1988 ؛Conner, 1977)؛ اما از لحاظ نظری، فرایند انتخاب کاملاً اندازه‌گیری شده و شناخته‌شده است. به خاطر این نکته کلیدی است که روشهای ساده آماری مانند تحلیل کواریانس (ANCOVA) می‌تواند برآوردهایی بدون‌سوگیری از اثر مداخله در طرحهای RD ارائه کنند (Overall & Woodward, 1977). در دیگر شبه‌آزمایش‌ها، انتخاب بین شرایط نه کاملاً شناخته‌شده است، و نه به درستی قابل‌اندازه‌گیری

[406] Probabilistically

(Overall & Woodward, 1977 ؛Lord, 1967). در بهترین حالت، متغیر تخصیص عبارتست از یک یا چند متغیر، که به طور جزئی مشاهده‌شده، و با خطا مورد اندازه‌گیری قرار گرفته‌اند.

این حالت را با جزئیات بیشتر در نظر بگیرید. تصور کنید که شرکت‌کنندگان را بر اساس نمره برش در آزمون IQ (مثلاً ۱۳۰) به گروهها تخصیص می‌دهید. این نمره فقط یک عدد است. افراد معمولاً از نمره برای استنباط در خصوص یک سازه استفاده می‌کنند؛ و معمولاً درباره اینکه این نمره، با چه دقتی سازه را اندازه‌گیری می‌کند، توافق ندارند. برخی ادعا می‌کنند که IQ هوش را اندازه‌گیری می‌کند، اما برخی دیگر می‌گویند این آزمون مواجهه با فرصت‌های آموزشی را اندازه‌گیری می‌کند. برای هر کدام از این استنباط‌ها، نمرات IQ واجد خطاست (نمره IQ برابر با ۱۳۰ کسب‌شده در یک آزمایش معین، معیار کاملی از هوش یا فرصت‌های شما نیست). اما در طرح RD، نمرات IQ برای اندازه‌گیری هوش یا فرصت به کار نمی‌روند. آنها تنها نشان می‌دهند که شرکت‌کنندگان باید به کدام گروه تخصیص داده شوند، و وقتی قرار باشد این نمرات تنها به عنوان مبنایی برای تخصیص به گروهها به کار گرفته شوند، واجد هیچگونه خطایی هم نیستند.

در دیگر شبه‌آزمایش‌ها، تخصیص با استفاده از سازه‌ای انجام می‌شود، که تنها با سطحی از خطا قابل اندازه‌گیری است. برای نشان دادن این مسئله، شکل ۷.۸ سه نمودار پراکندگی ارائه می‌کند که در آنها نمرات پیش‌آزمون روی خط افقی، و نمرات پس‌آزمون در همان متغیر روی محور عمودی قرار گفته‌اند، و نمرات پیش‌آزمون همبستگی کاملی با نمرات پس‌آزمون دارند (بحث را می‌توان به متغیر تخصیصی به غیر از پیش‌آزمون، و متغیرهایی با همبستگی ناکامل نیز تعمیم داد). در هر نمودار، نقاط موجود در سمت راست، مروبط به گروه مداخله هستند (N=200)، و نقاط واقع در سمت چپ به گروه کنترل تعلق دارند (N=200). در هر سه نمودار، مداخله تأثیری ندارد، اما خطای اندازه‌گیری تصادفی، به شکل سازنده‌ای تغییر می‌کند. در شکل ۷.۸ (الف)، نه پیش‌آزمون، و نه پس‌آزمون چنین خطایی ندارند. بنابراین، محل برخورد نمرات پیش‌آزمون و پس‌آزمون برای هر شرکت‌کننده بر روی یک خط مستقیم قطری قرار می‌گیرد، زیرا نمرات پس‌آزمونی که به صورت کامل اندازه‌گیری شده باشند، برابر با نمرات پیش‌آزمون هستند. این خط، مبنایی است که بر اساس آن می‌توانیم اثرات خطای اندازه‌گیری را برآورد کنیم.

در شکل ۷.۸ (ب) - همانطور که در غالب پیش‌آزمونها رایج است- خطا به پیش‌آزمون افزوده می‌شود. هر نقطه‌ای که پیشتر در شکل ۷.۸(الف) روی خط قطری قرار داشت، اکنون به شکل افقی جابه‌جا شده، و به صورت تصادفی در سمت راست یا چپ آن خط قرار گرفته است. به دلیل رگرسیون به میانگین ناشی از خطای اندازه‌گیری، هیچ‌کدام از شیب‌ها در شکل ۷.۸(ب) به تندی شیب ۷.۸(الف) نیست. برای مثال، نمرات پیش‌آزمون گروه مداخله که برابر بود با ۶۳، در پس‌آزمون به ۶۰/۹۹ کاهش می‌یابد؛ در حالیکه نمرات پیش‌آزمونی ۵۷ به ۵۸/۰۹ افزایش می‌یابد. به عبارت دیگر، خطای اندازه‌گیری در پیش‌آزمون سبب می‌شود که خط رگرسیون نمرات مشاهده‌شده،

شیب خود را تغییر دهد، حتی اگر خط رگرسیون نمرات واقعی (در شکل ۷.۸الف) بدون‌تغییر باقی مانده باشد. اگر خط رگرسیون را به دو گروه، به طور جداگانه برازش کنیم (همانگونه که در شکل ۷.۸ب دیده می‌شود)، یک اثر مداخله مصنوعی- به صورت شکافی بین خطوط رگرسیون- مشاهده می‌کنیم، اگرچه می‌دانیم که چنین اثری در حقیقت وجود ندارد. این اثر نادرست، همان چیزی است که در طرحهای گروه مقایسه غیرهم‌ارز، در مواقعی که مقیاس تخصیص به گروه‌ها با خطا اندازه‌گیری شده باشد، رخ می‌دهد.

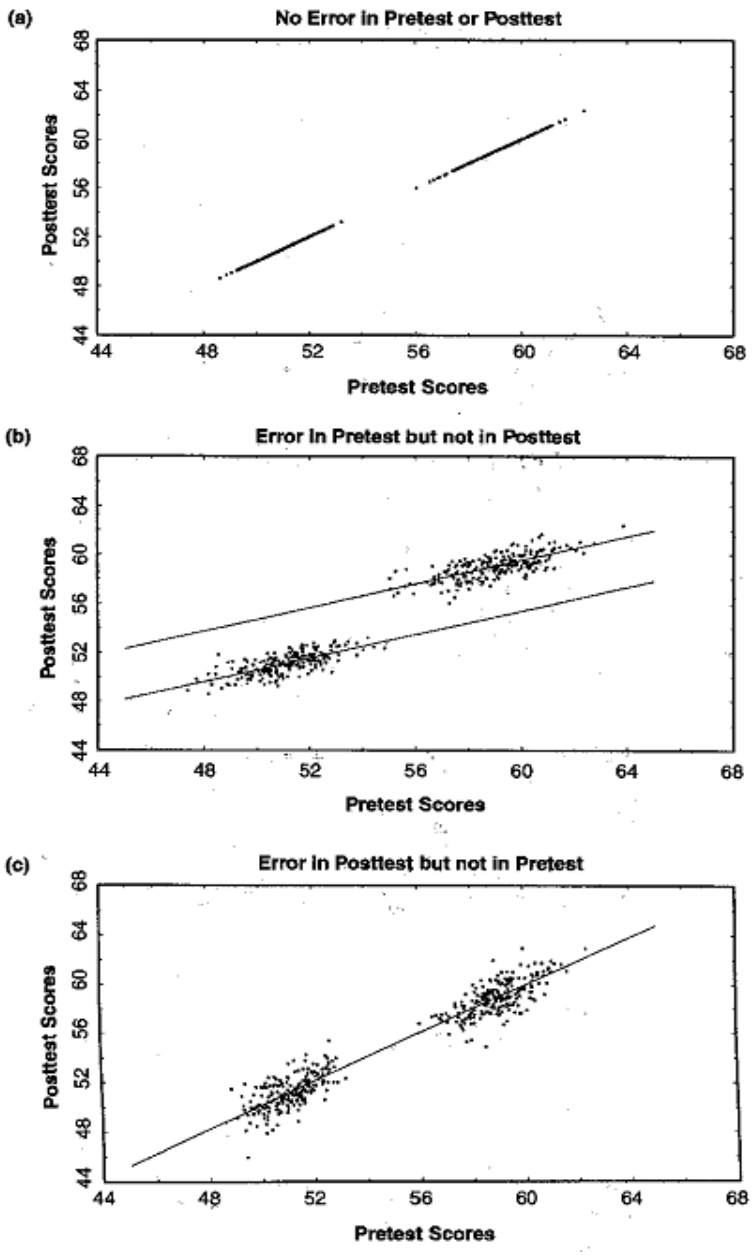
اگر خطا را به پس‌آزمون اضافه کنیم (شکل ۷.۸ج)، چنین سوگیری‌ای وجود نخواهد داشت. هر نقطه‌ای که پیشتر در شکل ۷.۸(الف) روی خط قطری قرار داشت، اکنون به صورت عمودی و تصادفی جابه‌جا شده، و در بالا یا پایین خط قرار گرفته است. اگر خطوط رگرسیون برای هر گروه به صورت جداگانه رسم شوند، خطوط هم‌پوشانی خواهند داشت، و همانگونه که در شکل ۷.۸(الف) دیده می‌شود، هیچ شکست یا ناپیوستگی‌ای وجود ندارد؛ که این حالت به درستی نبود اثر مداخله را نشان می‌دهد.[۴۰۷] شکل ۷.۸(ج)، دقیقاً موقعیتی در طرح ناپیوستگی رگرسیون که در آن مداخله اثری ندارد را نشان می‌دهد (به شباهت میان شکل ۷.۸ج و شکل ۷.۱ دقت کنید). در طرحهای RD، پس‌آزمون بدون شکل واجد خطاست، اما پیش‌آزمون (در این مورد، متغیر تخصیص) زمانی که به عنوان مقیاسی برای تخصیص به مداخلهها بکار گرفته می‌شود، خطایی ندارد.

خطوط رگرسیون تحت‌تأثیر خطاهای پس‌آزمون نیستند، اما خطاهای پیش‌آزمون در آنها سوگیری ایجاد می‌کند؛ زیرا رگرسیون کمترین مربعات معمولی[۴۰۸] (OLS) خطاهای متغیری که باید پیش‌بینی شود (طبق تعریف روی محور عمود قرار دارد) را کمینه می‌کند. در طرح RD، متغیر پس‌آزمون (یعنی خروجی) پیش‌بینی می‌شود، بنابراین رگرسیون OLS، اختلاف مربعی میان نمرات پس‌آزمونی مشاهده شده، و پیش‌بینی شده را به حداقل می‌رساند. از نظر گرافیکی، خطاها را جهت عمودی کمینه می‌کند؛ بنابراین، خط رگرسیون در شکل ۷.۸ج، دقیقاً در جایی قرار می‌گیرد که نمرات صحیح شکل ۷.۸الف در آن واقع بود. هیچکدام از این موارد، برای خطاهای پیش‌آزمون صحیح نیستند، زیرا خطاها به صورت افقی به پیش‌آزمون اضافه می‌شوند، اما رگرسیون همچنان خطاهای مربعی را به صورت عمودی کمینه می‌کند. بررسی بصری ساده شکل ۷.۸ب نشان می‌دهد که خط رگرسیون حاصله، در جهت مشخص‌شده جابجا می‌شود، و این سبب شکل‌گیری اثر مداخله واجدسوگیری در نتایج می‌شود.

---

[۴۰۷] خطای اندازه‌گیری تصادفی در پس‌آزمون می‌تواند سبب ناپیوستگی رگرسیون تصادفی در یک مطالعه معین شود، اما مقدار انتظاری ناپیوستگی صفر است، بنابراین براورد همچنان ناااریب در نظر گرفته می‌شود.

[۴۰۸] Ordinary least squares



شکل ۷ – ۸. اثر خطاها بر پیش‌آزمون و پس‌آزمون

پایبندی به نقطه برش

یکی از محدودیت‌های عملی بسیار کلیدی در اجرای طرحهای ناپیوستگی رگرسیونی پایبندی به نقطه برش در هنگام تخصیص شرکت‌کنندگان به گروه‌ها است. مشکلات عدم‌پایبندی دلایل متعددی دارد.

تخطّی از نقطه برش

در طرحهای RD، تخصیص مداخله باید بر اساس نقطه برش انجام شود؛ و این معمولاً در تناقض با انتظارات حرفه‌ای‌های حوزه مداخله است، که فکر می‌کنند قضاوت آنها باید مبنای اینکه چه کسی مداخله را دریافت کند، قرار داشته باشد. اگر این قضاوت کمّی نبوده، و بخشی از متغیر تخصیص را شکل ندهد، استفاده از چنین استراتژی-های قضاوتی، موجب تخطّی از پیشفرضهای طرح می‌شود. به عنوان مثال، رابینسون، برادلی و استنلی (Robinson, Bradley, & Stanley, 1990) نه تنها بر اساس برش، بلکه بر اساس قضاوتهای کیفی یک کمیته، شرکت‌کنندگان را به گروه‌ها تخصیص دادند. همچنین، مدیران مدارس معمولاً مایلند قدرت تخطّی از مقرارت تخصیص مداخله را داشته باشند (Trochim, 1984) تا بتوانند آنچه مطلوب آنهاست را انجام دهند، و زمانی که قضاوتهای حرفه‌ای و نمرات آزمون برای تعیین کسانی که نیاز به مداخله دارند، در تناقض با یکدیگر قرار دارند، بتوانند نقطه برش را کنار بگذارند. زمانی که بر اساس قضاوت نیاز به مداخله وجود داشته باشد، اما نمرات آزمون دال بر وجود این نیاز نباشد، به آسانی نمی‌توان مدیران را وادار کرد تا تنها بر اساس نمرات عمل کنند. پذیرش افراد در یک برنامه بدون در نظر گرفتن نمره برش، سبب ایجاد سوگیری می‌شود. در صورت توجیه‌پذیر بودن، هر کدام از این موارد باید شناسایی شده و قبل از تخصیص از نمونه تحقیقاتی حذف شوند. بهتر است افراد بدون بررسی نمرات واجد شرایط بودن، و بدون دانستن اینکه فرد به کدام گروه تخصیص داده می‌شد، حذف گردند.

اگر نتوان مواردی از این دست را در ابتدای کار حذف کرد، می‌توان آنها را در تحلیل حفظ کرد البته باید آنها را بر اساس نمرات شایستگی‌شان طبقه‌بندی کرد، و نه بر اساس مداخلههایی که در واقع دریافت کرده‌اند. این کار باعث می‌شود برآورد اثری بدون سوگیری بدست بیاوریم؛ البته بیشتر برآوردی از اثرات تخصیص به مداخله تا اثر مداخله. همچنین می‌توان آزمون را هم با وجود این افراد، و هم بدون آنها انجام داد تا مشخص شود که چه تغییری در برآوردها حاصل می‌شود. تروخیم (Trochim, 1984) دریافت که کنار گذاشتن اینگونه شرکت‌کنندگان، برآورد اثرات را بسیار دقیق‌تر می‌کند، اما احتمالاً میزان اشتباه رخ داده در تخصیص مداخله، تفاوتهایی ایجاد می‌کند. با این وجود، اگر تخصیص نادرست، حاصل نقض عامدانه نمره تخصیص باشد، هیچ درمان تصحیحی روش‌شناختی وجود ندارد، اگرچه ممکن است بتوان جهت سوگیری احتمالی را برآورد کرد.

اگر شرکت‌کنندگان خیلی کند یا خیلی سریع وارد شوند، تخصیص مبتنی بر برش دشوار خواهد شد؛ و این مسأله نوعی فشار روی محقق ایجاد می‌کند تا نمره برش را تعدیل کند، به‌گونه‌ای که بتواند تعداد حداقل شرکت‌کننده لازم برای حفظ برنامه را داشته باشد (Havassey, 1988). اگر اندازه نمونه به اندازه کافی بزرگ باشد، گاهی می‌توان برش را تعدیل کرد، اما باید هر گروه را با استفاده از یک نقطه برش متفاوت، و در قالب یک طرح RD جداگانه تحلیل کرد. یا اگر تعداد شرکت‌کنندگان شایسته بیشتری از آنچه می‌توان مداخله روی آنها انجام داد داشته باشیم، می‌توان بخشی از افراد اضافی را به صورت تصادفی به گروه فاقدمداخله تخصیص داد، و یک طرح تصادفی را درون طرح ناپیوستگی رگرسیون پایه مستتر کرد.

اگر نمره برش عمومی باشد، شرکت‌کنندگان بالقوه، می‌توانند نمره خود را دستکاری کنند تا بتوانند در مداخله وارد شوند. در مطالعه RD انجام شده توسط برک و راما (Berk and Rauma, 1983) بر روی اثر پرداخت کمک هزینه بیکاری روی بازگشت به جرم زندانیان سابق، زندانیانی که نزدیک به نقطه برش بودند، ممکن است ساعت‌ها تلاش کنند تا بتوانند به مداخله وارد شوند، در حالیکه آنهایی که از پیش بدانند که در گروه وارد شده‌اند، یا شانس بالایی برای وارد شدن ندارند، تلاشی نخواهند کرد. این مسأله بر توزیع نمرات زندانیان در متغیر تخصیص تأثیرگذار خواهد بود، و در بدترین حالت، شکافی دقیقاً پایین نمره برش ایجاد می‌کند که می‌تواند منجر به شکل‌گیری داده‌های دومُدی[۴۰۹] شود. مثالی دیگر، آزمون ترک‌تحصیل‌کنندگان ایرلندی بود که در آن، امتیازدهندگان اکراه فراوانی نسبت به تخصیص نمرات کمّی پایین‌تر (دقیقاً پایینتر با فاصله اندک) از نقطه برش نشان دادند (Madaus & Greaney, 1985). داده‌هایی از این دست که به صورت غیر طبیعی توزیع می‌شوند، سبب ایجاد سطوح رگرسیون غیرخطی می‌گردند، که تحلیلها را پیچیده می‌کند.

تداخل[۴۱۰] و ریزش

تداخل مداخله زمانی رخ می‌دهد که افراد تخصیص‌یافته به گروه مداخله، در آن شرکت نکنند، یا کسانی که به گروه کنترل تخصیص داده شده‌اند، در نهایت سر از گروه مداخله دربیاورند. به عنوان مثال، رابینسون و استنلی (Robinson and Stanley, 1989) برخی از شرکت‌کنندگان که به مداخله تخصیص داده شده، اما در آن شرکت نکردنده بودند را در گروه کنترل وارد کردند. در فصل بعد، نشان خواهیم داد که چطور اینگونه تداخلات به طور بخشی در آزمایش‌های تصادفی لحاظ می‌شوند؛ برخی از آن اصول را می‌توان برای طرحهای RD نیز بکار برد.

مشکل دیگر، ریزشی است که بعد از تخصیص، در مطالعه رخ می‌دهد. سیور و کوارتون (Seaver & Quarton, 1976) اثرات اعطای تقدیرنامه به دانش‌آموزان بر اساس کسب معدل برابر با ۳.۵، یا بیشتر در ترم را مورد مطالعه قرار دادند. اما آنها هر دانشجویی که در پاییز، زمستان و بهار بعد ثبت نام نکرده بود، یا دانشکده خود را تغییر داده بود، یا کسانی که کمتر از نُه واحد در هر ترم برداشتند، را کنار گذاشتند (صفحه ۴۶۰). حذف‌های صورت‌گرفته با در نظر گرفتن تعریف جمعیت لازم بود. در اغلب تحقیقات میدانی، این فرسایش اجتناب‌ناپذیر است. اما کنار گذاشتن شرکت‌کنندگان بر اساس اندازه‌گیری‌هایی که می‌تواند ناشی از تخصیص به مداخله باشد، مشکل‌زاست. به عنوان مثال، اگر تقدیرنامه به دانشجویی این شانس را بدهد که وارد کالج بهتری شود، و اگر این امر دلیل انتقال او باشد، آنگاه می‌توان گفت که ریزش با مداخله همبستگی دارد، و در نتیجه، برآوردهایی دارای سوگیری خواهیم

[409] Bimodal
[410] Crossover

داشت. باز هم، راه‌حلهای بخشی (جزئی) که بعداً برای آزمایش‌های تصادفی ارائه خواهیم کرد، را می‌توان در طرحهای RD نیز بکار گرفت.

ناپیوستگی رگرسیونی فازی

تروخیم (Trochim, 1984) مطالعاتی که در آنها، تخصیص به گروه مداخله، با پایبندی کامل به نقطه برش صورت نمی‌گیرد (مثلاً Cahan & Davis, 1987) را طرح ناپیوستگی رگرسیونی فازی می‌نامد. اگر بخواهیم صریح باشیم باید گفت که طرح RD با نقطه برش فازی، اصلاً طرح RD نیست. بلکه یک آزمایش تصادفی تنزل‌یافته است که همچنان می‌تواند برآوردهای بهتری نسبت به انواع طرحهای شبه‌آزمایشی تولید کند (Shadish & Ragsdale, 1996)؛ البته به شرطی که فازی بودن خیلی شدید نباشد.

اگر محدوده تخصیص نادرست حول نمره برش به یک دامنه باریک محدود باشد، یعنی مثلاً در شکل ۷.۲، بین نمرات ۴۹/۵ و ۵۰/۵ برای متغیر تخصیص قرار گیرد، آنگاه شرکت‌کنندگان در این محدوده را می‌توان حذف کرد و باقیمانده شرکت‌کنندگان را به عنوان طرح RD محدود در نظر گرفت (Mohr, 1988, 1995). این راه حل، تنها زمانی خوب کار می‌کند که دامنه‌ای که باید حذف شود، باریک باشد و در غیر اینصورت مدلسازی دقیق خط رگرسیون در نزدیکی برش، کار دشواری است. اگر کمتر از ۵٪ شرکت‌کنندگان به اشتباه تخصیص داده شده باشند، می‌توان آنها را کنار گذاشته، و برآوردهای اثر مداخله معقول‌تری بدست آورد (Judd & Kenny, 1981a؛ Trochim, 1984). برک و دیلاو (Berk and deLeeuw, 1999) برای دیگر موارد که در آنها درصد شرکت‌کنندگان با تخصیص نادرست، بالاتر از ۵٪ است، یا در محدوده گسترده‌تری از متغیر تخصیص قرار دارند، تحلیل‌های حساسیت سودمندی ارائه کرده‌اند، که با کمک آنها می‌توان آثار انحراف از فرایند تخصیص روی خروجی را بررسی کرد. در نهایت، در چنین حالت‌هایی با کاربرد مدل‌های سوگیری انتخاب، یا تحلیل‌های نمره تمایل که در فصل ۵ مورد بحث قرار گرفت، می‌توان برآوردها را بهبود داد. اگرچه تحقیقاتی در این حوزه صورت نگرفته است.

تهدیدهای روایی

منطق آماری در پس طرح RD سبب شده تا برای مواقعی که تخصیص تصادفی امکان‌پذیر نیست، این طرح جایگاه ویژه‌ای در مطالعات علّت‌یابی داشته باشد. اما قدرت این طرح زمانی روشنتر می‌شود که آن را با سری زمانی مختل شده[411] (ITS) -شبه‌ازمایشی که از نظر مفهومی شبیه‌ترین طرح به آنهاست- مقایسه کنیم.

ناپیوستگی رگرسیون و سری زمانی مختل شده

[411] Interrupted time series

در هر دو طرح سری‌زمانی مختل‌شده و طرح RD، پیش‌بینی می‌شود اثر در نقطه خاصی از یک طیف رخ دهد. در طرح سری‌زمانی مختل‌شده، آن طیف زمان است، و در طرح RD، آن طیف عبارتست از متغیر تخصیص. در طرح سری‌زمانی مختل‌شده، مداخله در نقطه زمانی از پیش‌معینی رخ می‌دهد، و در طرح RD، در نمره برش مشخصی رخ می‌دهد. اگر مداخله در طرح سری‌زمانی مختل‌شده مؤثر باشد، میانگین یا شیب سری‌زمانی در نقطه‌ای که مداخله اعمال‌شده، تغییر خواهد کرد؛ و در طرح RD، یک مداخله مؤثر می‌تواند شیب یا عرض از مبداء خط رگرسیون را در نقطه برش معین تغییر دهد. با وجود اینکه نمودارهای طرح سری‌زمانی مختل‌شده و طرح RD مشابه یکدیگر به نظر می‌رسند.

بنابراین تعجبی ندارد که تهدیدهای روایی در طرح RD، تفاوت اندکی با عوامل تهدیدکننده طرح سری‌زمانی مختل‌شده داشته باشد. در سری‌زمانی مختل‌شده ساده، مهم‌ترین تهدیدها، مختص نقاط هستند، و همزمان با زمان اجرای مداخله رخ می‌دهند. معمولاً شکل‌های ساده بلوغ یا انتخاب به ندرت منجر به تهدید می‌شوند؛ اما گذشت زمان (رویدادهای خارجی که همزمان با مداخله رخ داده‌اند)، و تغییر تجهیزات در زمان مداخله، می‌تواند گاهی تهدیدآمیز باشد. خواهیم دید که این حالت برای طرحهای RD نیز صادق است. همچنین در سری‌زمانی مختل‌شده برای تخمین اثر نیازمند آن هستیم که خودهمبستگی، روند، چرخه‌ها و رانش[۴۱۲] را به درستی مدلسازی کنیم؛ یعنی باید شکل طبیعی سری‌زمانی را بدانیم تا بتوانیم تغییرات این شکل را شناسایی کنیم. تحلیل آماری RD نیز به همین ترتیب پیچیده است.

اعتبار نتایج آماری[۴۱۳] و تعیین نادرست[۴۱۴] شکل تابع

وقتی که آثار مداخله در سری‌زمانی مختل‌شده بزرگ باشند، و همزمان با آغاز اجرای مداخله رخ دهد، اغلب اثر بدون نیاز به انجام تحلیل آماری قابل پذیرش و موجّه است. در طرحهای RD نیز، وقتی که ناپیوستگی در نقطه برش بزرگ باشد (مانند مثال پزشکی شکل ۷.۴)، به تحلیل آماری اندکی نیاز است. با این وجود، این آثار شدید به ندرت در طرحهای RD یا سری‌زمانی مختل‌شده دیده می‌شوند. بنابراین، مدلسازی صحیح شکل [خط] رگرسیون برای تشخیص اثرات RD ضروریست. در ساده‌ترین حالت، که در آن خط رگرسیون خطی است، تحلیل کواریانس (ANCOVA) مانند معادله زیر مورد استفاده قرار می گیرد:

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 Z_i + \hat{\beta}_2 \left(X_i - X_c \right) + e_i \qquad (۷.۱)$$

412 Drift
413 Statistical conclusion validity
414 Misspecification

در این معادله، Y خروجی است، $\hat{\beta}_0$ [۴۱۵] عرض از مبداء است، Z متغیر دومقداره (۰،۱) است که نشان می‌دهد، شرکت‌کننده کدامیک از شرایط مداخله را دریافت می‌کند (۰ = کنترل، ۱ = مداخله)، X متغیر تخصیص است، ضریب رگرسیون پیش‌بینی‌کننده خروجی از روی متغیر تخصیص، $\hat{\beta}_2$ است، اثر مداخله با ضریب رگرسیون $\hat{\beta}_1$ اندازه‌گیری می‌شود، و e یک جمله خطای تصادفی است. اندیس i، نشان‌دهنده N واحد درون مطالعه است (از i=1 تا N). تفریق مقدار برش از متغیر تخصیص (X_i-X_c ، که مشابه در مرکز قرار دادن متغیر تخصیص است، زمانی که برش برابر با میانگین باشد) باعث می‌شود تا معادله بتواند اثر مداخله را در نمره برش- یعنی در نقطه‌ای که گروه‌ها در آن بیشترین شباهت را دارند- برآورد کند. می‌توان با تغییر مقداری که باید تفریق شود، اثر را در هر جایی از محدوده متغیر تخصیص برآورد کرد یا با تفریق از صفر، آن را روی عرض از مبداء بدست آورد.

این تحلیل ساده، برآوردی بدون‌سوگیری از اندازه اثر مداخله بدست می‌دهد. اثبات این مسئله آماری بوده، و در اینجا از آن صرفنظر می‌کنیم (ضمیمه۷.۱ را ببینید). درک این اثبات‌ها، روشن می‌سازد چطور طرحهای RD به آزمایش‌های تصادفی و مدلهای سوگریی انتخاب (که در ضمیمه ۵.۱ به آنها اشاره شد) شباهت دارند. با این وجود، براورد اثر مداخله در طرحهای RD تنها زمانی بدون‌سوگیری خواهد بود که مدل ساده (۷.۱) صحیح باشد. دو مشکل وجود دارد که می‌توانند سبب نادرستی مدل شوند: غیرخطی‌بودن و برهم‌کنش‌ها.

غیرخطی بودن

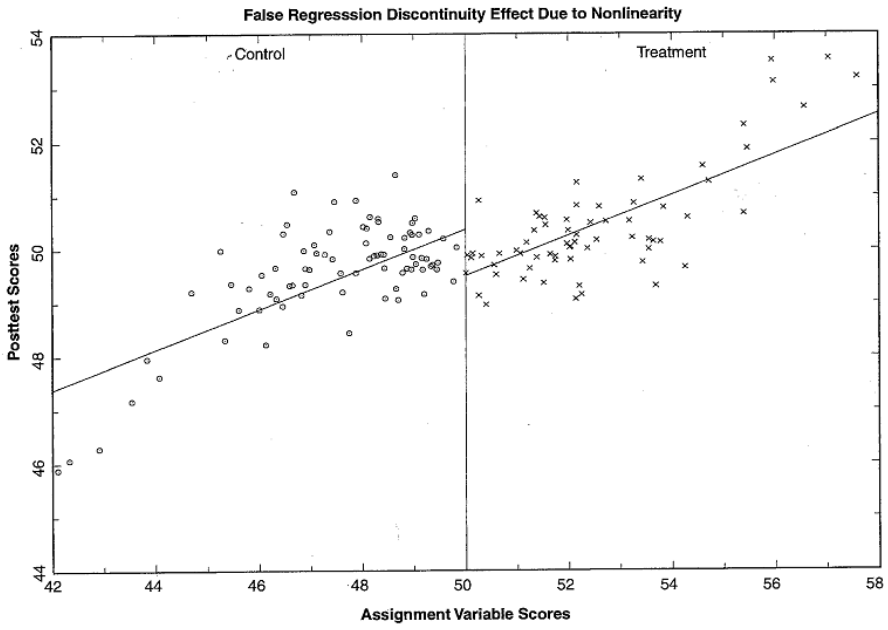
مدل (۷.۱) نشان می‌دهد که رابطه بین متغیر انتخاب و متغیر نتیجه‌ای خطی است. فرض کنید که این رابطه در واقع غیرخطی باشد، و مثلاً به جای تنها X ، تابعی از X^2 یا X^3 هم باشد. اگر مدل، این جملات غیرخطی را لحاظ نکرده باشد، آنگاه مدل به طور نادرست تعیین‌شده[۴۱۶]، و برآوردهای ناپیوستگی رگرسیون ممکن است مانند آنچه در شکل ۷.۹ دیده می‌شود، سوگیرانه باشند. در این شکل، داده‌ها به گونه‌ای ساختار یافته‌اند که پس‌آزمون، تابعی مکعبی از متغیر تخصیص است (تابعی از X^3) و مداخله در حقیقت اثری نداشته است. اگر داده‌ها به جای اینکه به مدلی با جمله معکبی برازش شوند، به غلط به معادله ۷.۱ برازش شوند، ناپیوستگی‌ای بزرگ اما غلط، بین دو خط رگرسیون در محل نقطه برش ظاهر می‌شود. در این مورد، ممکن است به اشتباه گفته شود که مداخله اثر منفی قابل‌توجهی داشته است. برای بدست آوردن جواب صحیح، باید معادله جمله X^3 داشته باشد. کوک و کمپبل (Cook & Campbell, 1979) مثالی از تحلیل مجدد خود روی آزمایش RD سیور و کورتون (Seaver and Quarton,

[۴۱۵] جود و کنی (Judd and Kenny, 1981a)، احتمالات مشکلات تحلیل برای بیش از یک متغیر خروجی را نشان دادند و برک و راما (Berk and Rauma, 1983) تحلیل را برای یک خروجی دو بخشی انجام دادند.

[416] Misspecified

1973) انجام دادند. آنها در ابتدا اثر ناشی از حضور در فهرست رؤسا مشاهده کردند، اما زمانی که مدل منحنی‌شکل به جای مدل خطی جایگزین شد، این اثر غیرمعنادارشد.

اگر متغیرها به صورت نرمال توزیع نشوند نیز، حالت غیرخطی ایجاد می‌شود. تروخیم، کاپلری و ریچارد (Trochim, Cappelleri & Reichardt, 1991) در شکل‌های ۱ و ۲ این حالت را با استفاده از داده‌هایی که به طور یک‌شکل (یعنی در هر نقطه از پیش‌آزمون، تعداد تقریباً برابری از واحدها وجود داشتند)، و نه نرمال توزیع شده بودند نشان می‌دهند. انحنای[۴۱۷] حاصله، به شکل ۷.۹ شباهت داشت و یک شبه‌اثر[۴۱۸] از برازش داده‌ها به مدل خطی بدست آمد. بررسی توزیع متغیرهای تخصیص و متغیرهای پس‌آزمون برای شناسایی این رابطه مفید است؛ تبدیل[۴۱۹] داده‌ها گاهی به حل این مشکلات توزیع کمک می‌کند (Trochim, 1984). غیرخطی بودن می‌تواند ناشی از نقاط دورافتاده شانسی یا آثار سقف و کف باشد. خوشبختانه در همه این موارد، باز هم تابع مکعبی روی داده‌ها برازش می‌شود. که این نشان می‌دهد که اثری وجود نداشته، و ناپیوستگی (مشاهده شده در بررس چشمی نمودار) در واقع وجود ندارد.



شکل ۷ – ۹ اثر ناپیوستگی رگرسیون نادرست به علت غیرخطی بودن

[417] Curvilinearity
[418] Pseudoeffect
[419] Transformation

برهم‌کنش‌ها

غیرخطی‌بودن همچنین می‌تواند ناشی از ناتوانی در مدل کردن آماری برهم‌کنش‌های موجود میان متغیرهای مداخله و تخصیص باشد. معادله (۷.۱)، تنها یک جمله اصلی مربوط به اثر مداخله را تعیین می‌کند، اما جملات برهم‌کنشی را در برنمی‌گیرد. اگر واحدهایی که به گروه مداخله تخصیص می‌یابند و امتیازی نزدیک به نقطه برش دارند، کمتر از گروه‌هایی که امتیازات حدی‌تر دارند، بهره‌مند شوند چه؟ در این مورد، باید یک جمله حاصلضرب[۴۲۰] برای برهم‌کنش به معادله افزوده شود تا بتوان اثری بدون سوگیری بدست آورد[۴۲۱]. اگر داده‌ها نه بر اثر مداخله اصلی، که به دلیل وجود یک اثر برهم‌کنش تولید شده باشند، و اگر جمله برهم‌کنش مورد نظر [در معادله] گنجانده نشده باشد، ضریب بدست‌آمده برای اثر اصلی سوگیرانه خواهد بود، که به شکل ناپیوستگی نادرست در محل نقطه برش (در شکل ۷.۱۰) نمایان خواهد شد. در این شکل، ضریب مربوط به Z_i باید صفر باشد، اما در عوض می بینیم که در محل $\hat{\beta}_1 = 1.42$ معنادار است؛ دیگر ضرایب این معادله نیز سوگیرانه خواهند بود.

شکل ۷.۱۱، یک مورد مدلسازی شده صحیح که دربرگیرنده اثر اصلی مداخله و برهم‌کنش بین نمرات پیش‌آزمون و مداخله است را نشان می‌دهد. علاوه بر ناپیوستگی رگرسیونی در محل نقطه برش، شیب خط نیز در سمت راست برش، تندتر از سمت چپ است. در این مثال، همه شرکت‌کنندگان گروه مداخله بیش از افراد گروه کنترل منتفع می‌شوند، اما آنهایی که نمرات بالاتری در متغیر تخصیص دارند، عملکرد بهتری نسبت به افرادی دارند که نمرات پایین‌تری در متغیر تخصیص کسب کرده اند. بنابراین، اندازه ناپیوستگی بسته به اینکه کجا بر روی متغیر تخصیص اندازه‌گیری شده باشد، تغییر می‌کند. بسته به نمره‌ای که از متغیر تخصیص تفریق می‌شود، این تحلیل براوردهای متفاوت اما صحیحی از اثر بدست می‌دهد.

شکل ۷.۱۲، تغییر در شیب را در محل نقطه برش نشان می‌دهد، بی‌آنکه ناپیوستگی وجود داشته باشد. این تغییر ناشی از برهم‌کنش است، نه اثر مداخله. در این حالت، برخی به علّت عدم‌وجود ناپیوستگی در محل برش می‌گویند اثری وجود ندارد، اما برخی دیگر آن را نشانه یک اثر احتمالی در محلی دور از نقطه برش می‌دانند. تفسیر دوم با دو مشکل مواجه است. اول اینکه منطق طرح RD تا حدی به یافتن ناپیوستگی‌ها در نقطه برش متکی است، زیرا

---

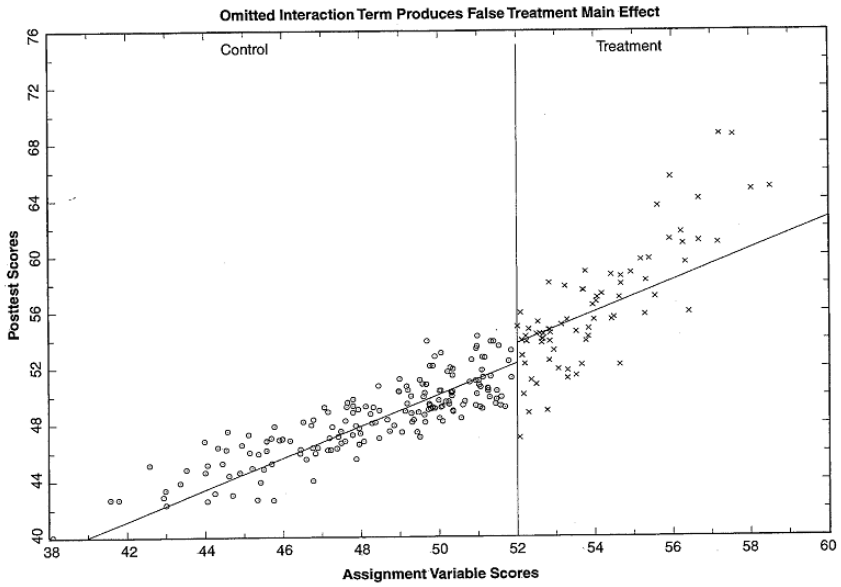
[420] Product term

[۴۲۱] این جمله در حالتی که نمره برش از متغیر تخصیص X تفریق شود، به صورت $Z_i\left(X_i - X_c\right)$ خواهد بود. بنابراین، معادله باید برای مدلسازی حالت غیرخطی و برهم‌کنش‌ها از توابع متغیر تخصیص بهره بگیرد که درجه چندجمله‌ای را افزایش می‌دهند $X, X^2, X^3, ..., X^n$ و همچنین یک جمله برهم‌کنشی برای هر کدام از چندجمله‌ای‌ها داشته باشد ($ZX, ZX^2, ZX^3, ..., ZX^n$):

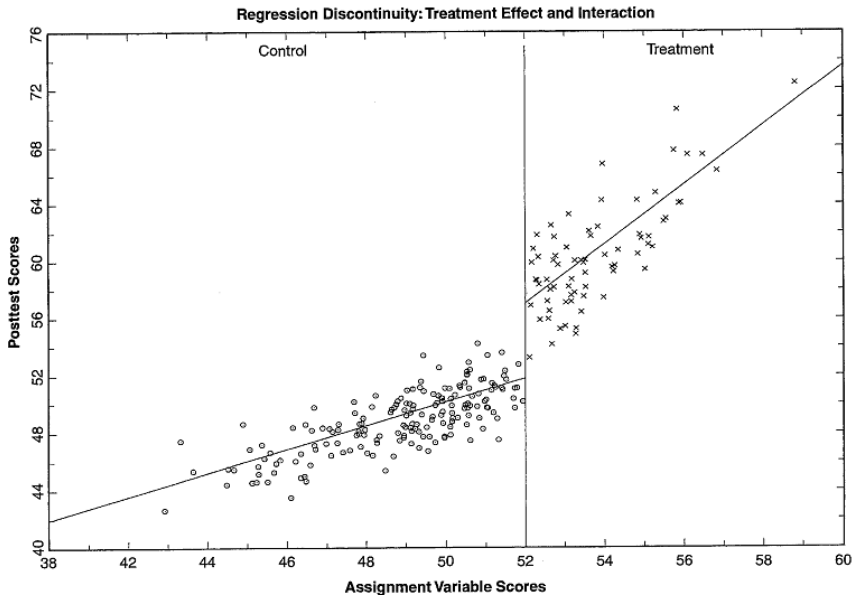
$$Y_i = b_o + \hat{\beta}_i Z_i + \hat{\beta}_2\left(X_i - X_c\right) + \hat{\beta}_3 Z\left(X_i - X_c\right) + \hat{\beta}_4\left(X_i - X_c\right)^2 + \hat{\beta}_5 Z_i\left(X_i - X_c\right)^2 + ... + \hat{\beta}_{n-1}\left(X_i - X_c\right)^S + \hat{\beta}_n Z_i\left(X_i - X_c\right)^S + e_i \quad (۲-۷)$$

که جملات آن پیشتر توصیف شده‌اند و S در جه بالاترین چندجمله‌ای برازش شده به مدل است و $\hat{\beta}_n$ ضریب رگرسیون آخرین جمله برهم‌کنشی یا چندجمله‌ای در مدل است.

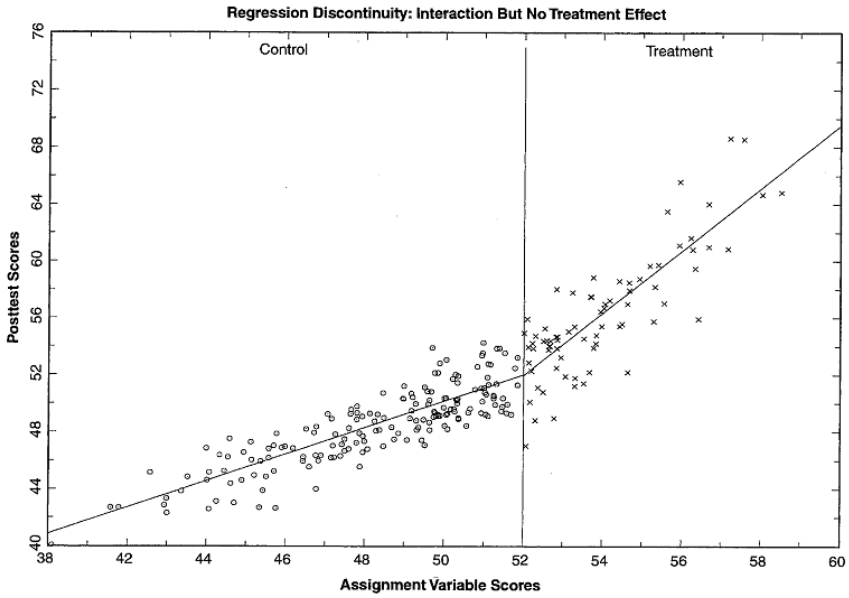
شرکت‌کنندگان در آن نقطه بیشترین شباهت را دارند. دوم اینکه، به سختی می توان تمایزی میان شکل ۷.۱۲ و رابطه غیرخطی حاصل از تابعی مربعی در حالتی که مداخله بی‌اثر بوده است، مشاهده کرد. گاهی می‌توان این احتمالات را برای ترکیب طرح RD با دیگر مؤلفه‌های طراحی (که در ادامه به آنها خواهیم پرداخت) بکار گرفت.



شکل ۷.۱۰. جمله برهم‌کنشی حذف شده، سبب ایجاد اثر مداخله نادرست می شود



شکل ۷.۱۱. ناپیوستگی رگرسیونی: اثر مداخله و اثر برهم‌کنش



شکل ۷.۱۲. ناپیوستگی رگرسیونی: اثر برهم‌کنش، بدون وجود اثر مداخله

توصیه‌هایی در باب مدلسازی شکل تابعی

اگر احتمال وجود برهم‌کنش‌ها یا غیرخطی بودن وجود داشته باشد، تحلیلها باید مدل را بیش‌برازش[۴۲۲] کنند. برای این‌کار، تحلیل را با لحاظ کردن جملات برهم‌کنشی و چندجمله‌ای بیش از حد نیاز شروع کرده، سپس جملات غیرمعنادار را از مرتبه بالا به پایین حذف می‌کنیم. اگر اطمینان ندارید، این جملات را از معادله حذف نکرده و حفظ کنید. این بیش‌برازش، ضرایبی بدون سوگیری بدست می‌دهد، اگرچه توان را کاهش می‌دهد (Cappelleri & Trochim, 1992). محقق باید این موضوع را بررسی کند که آیا نتایج بدست‌آمده برای پیش‌فرضهای متفاوت درباره شکل تابعی، برقرار (پایدار) هستند یا نه. علّت‌یابی‌ها و اصلاحات استاندارد رگرسیونی نیز (مثلاً J. (West et al., 2000 ؛Neter, Wasserman & Kutner, 1983 ؛Cook & Weisberg, 1994 ؛Cohen, & Cohen, 1983 می‌تواند به تشخیص این مشکلات کمک کند. برای ارائه یک شکل تابعی پایه، اغلب می‌توان داده‌های اضافی که متأثر از مداخله نیستند -مانند داده‌های متغیرهای خروجی و متغیر تخصیص پیش از اجرای مداخله، و یا داده از متغیرهای تخصیص و خروجی بدست‌آمده از نمونه‌های مرتبط اما دستکاری‌نشده از شرکت‌کنندگان- جمع‌آوری کرد. اگر این کار ممکن نباشد، اما اندازه نمونه بزرگ باشد، می‌توان نمونه را دو نیم کرد؛ و با نیمه اول، مدل را

[422] Overfit

ساخته، و با نیمه دوم، آن را اعتبارسنجی[۴۲۳] کرد. در نهایت، برای رفع این مسأله می‌توان طرح‌های RD را با آزمایشهای تصادفی ترکیب کرد. در ادامه به توضیح چگونگی انجام اینکار خواهیم پرداخت، اما پیش از آن عملکرد طرحهای RD را از نظر تهدیدهای روایی درونی بررسی می‌کنیم.

اعتبار درونی

تهدید اعتبار درونی در RD، شکستی ناگهانی در خط رگرسیون، دقیقاً در محل نقطه برش، ایجاد می‌نماید. این امر تقریباً همیشه غیرموجّه است. در واقع، به استثنای نمودارهای پراکندگی ارائه شده در این فصل، اغلب خوانندگان هرگز نمودار پراکندگی‌ای را نخواهند دید که در آن یک شکست یا ناپیوستگی به طور طبیعی وجود داشته باشد. برای ایجاد چنین شکستی، تقریباً همواره باید یک اثر مداخله وجود داشته باشد که تنها بر افراد موجود در یک طرف نقطه برش اعمال شده باشد.

تهدید *انتخاب* نمی‌تواند چنین شکستی ایجاد کند. این امر به این دلیل است که ما می‌توانیم تهدید انتخاب را با موفقیت در مدل لحاظ کنیم؛ با در نظر گرفتن اینکه [جریان انتخاب] کاملاً شناخته شده بوده، و مورد اندازه‌گیری قرار گرفته است. اما حتی اگر مسائل آماری مرتبط با این بحث برای خواننده شفاف نباشند، می‌توان با این دلیل معقول آن را توجیه کرد که، اگر شرکت‌کنندگانی با نمرات تخصیص۵۰/۰۵ در شکل ۷.۱، از نظر متغیر خروجی بسیار بهتر از شرکت‌کنندگانی که نمره ۴۹/۹۵ (در متغیر تخصیص داشته‌اند) عمل کنند، مطمئناً اختلاف ۰/۰۵ نمره ای آنها در متغیر تخصیص نمی‌توانسته مسئول این تفاوت در عملکرد باشد. تهدید *گذشت زمان* در صورتی موجّه خواهد بود که رویدادهایی که می‌توانسته بر خروجی تأثیر داشته باشد، تنها روی افراد حاضر در یک طرف برش رخ داده باشند. که این اتفاق که به طور نمونه در مثال میدکید رخ داد (شکل ۷.۴)، معمولاً هم نامحتمل است و هم به آسانی قابل ارزیابی. تهدید *آزمون* نیز نمی‌تواند بر اندازه ناپیوستگی تأثیرگذار باشد، زیرا هر دو گروه، آزمایش‌های یکسانی دریافت می‌کنند. و تغییر در تجهیزات و ابزار آزمایش که دقیقاً در نقطه برش رخ داده باشد نیز، باز هم غیرمحتمل و به آسانی قابل ارزیابی است. تهدید *بلوغ* باید دلالت بر این داشته باشد که متغیر خروجی به صورت طبیعی برای آنهایی که متغیر تخصیص بالاتر دارند (نسبت به آنهایی که نمره کمتری در متغیر تخصیص گرفته اند)، سریع‌تر رشد کرده باشد؛ این امر می‌تواند سبب نوعی انحنا شود که نیازمند مدلسازی دقیق خواهد بود. دقیقاً همانند آزمایش‌های تصادفی، *مرگ‌ومیر* همواره می‌تواند تهدیدی برای طرحهای RD بشمار بیاید، البته اگر با تخصیص مداخله همبستگی داشته باشد (Shapiro, 1984).

ممکن است بسیاری از خوانندگان فکر کنند که رگرسیون آماری، در این طرحها تهدیدی موجّه است، زیرا گروهها بر اساس مقادیر حدّی در توزیع متغیر تخصیص، انتخاب و تشکیل می‌شوند. اما رگرسیون پیشتر به طور کامل در

[423] Cross-validate

معادله خط رگرسیون میان متغیر تخصیص و پس‌آزمون لحاظ شده است. ضریب همبستگی r، بین این دو آزمون، میزان رگرسیونی که رخ خواهد داد را اندازه‌گیری می‌کند. در واقع، هم جمله رگرسیون و هم نماد r به این دلیل بکار گرفته می‌شوند که رگرسیون به میانگین را اندازه‌گیری کنند. این درست است که، فردی که نمره بالایی در متغیر تخصیص دارد، در متغیر خروجی نمره بالایی نخواهد داشت، و فردی که نمره پایینی در متغیر تخصیص دارد، نمره‌ای به همان اندازه پایین در متغیر خروجی کسب نخواهد کرد. اما این مسأله تنها سبب افقی‌تر شدن خط رگرسیون می‌شود، و نمی‌تواند ناپیوستگی یا شکستی در نقطه برش ایجاد کند.

دیگر انواع تهدیدها نیز می‌توانند در این طرحها مسأله‌ساز باشند. به عنوان مثال، زمانی که اثر سقف یا کف در اندازه‌گیری‌ها وجود داشته باشد، ممکن است اثر ترکیبی انتخاب-ابزار[۴۲۴] رخ دهد. کاپلری و تروخیم (Cappelleri & Trochim, 1992) در مطالعه خود روی اثر داروی زاناکس بر اضطراب، اثر کف را اینگونه توصیف می‌کنند: دارو آنچنان موثر بود که در پایان مطالعه، در بسیاری از شرکت‌کنندگان گروه مداخله، هیچ نشانه‌ای [از اضطراب] باقی نمانده بود، و این باعث مسطح شدن رگرسیون در آن سطح کف شد. این امر سبب می‌شود تا در صورت وجود مدلسازی نادرست، طرحهای RD براوردهای سوگیرانه از اثر برهم‌کنش بدست دهند. با این وجود، در کل، طرحهای RD آسیب پذیری اندکی نسبت به تهدیدهای روایی درونی دارند.

ترکیب ناپیوستگی رگرسیونی و آزمایش‌های تصادفی

در صورت توجیه‌پذیر بودن، ترکیب آزمایش تصادفی با طرحهای RD، بسیار بهتر از کاربرد طرح RD به تنهایی است. به عنوان مثال، به جای استفاده از یک نمره برش، می‌توان یک بازه برش مابین دو نمره را تعریف کرد. نمرات بالاتر از حد بالا، همه شرکت‌کنندگان به یک شرایط تخصیص داده می‌شوند، و همه شرکت‌کنندگان دارای نمرات پایین‌تر از حد پایین نیز به شرایط دیگر. شرکت‌کنندگانی که نمرات قرار گرفته درون بازه را کسب کرده‌اند، به صورت تصادفی به یکی از شرایط موجود تخصیص داده می‌شوند. طرح حاصله، یک آزمایش تصادفی تعبیه شده (مستتر) در یک طرح ناپیوستگی رگرسیون است. به عنوان مثال در شکل ۷.۱۳، شرکت‌کنندگان با نمره تخصیص کمتر از ۴۸، به شرایط کنترل، افراد با نمره تخصیص بالاتر از ۵۲ به شرایط مداخله، و افراد دارای نمرات بین ۴۸ و ۵۲، به طور تصادفی به یکی از شرایط (کنترل یا مداخله) تخصیص داده شده‌اند. همه شرکت‌کنندگان در معادله (۷.۱) تحلیل شده‌اند، اما باید تعیین کرد که کدام نمره برش از نمره متغیر تخصیص کم شود. تروخیم (Trochim, 1991) پیشنهاد می‌کند از برشی استفاده شود که همه شرکت‌کنندگان گروه مداخله را به یک طرف برش تخصیص دهد. منطق او این است که برهم‌کنش بین تخصیص و مداخله می‌تواند بر همه شرکت‌کنندگان مورد مداخله، تأثیر بگذارد، از جمله آنهایی که درون بازه هستند؛ بنابراین تفریق آن برش کمک می‌کند تا بتوانیم این برهم‌کنش را

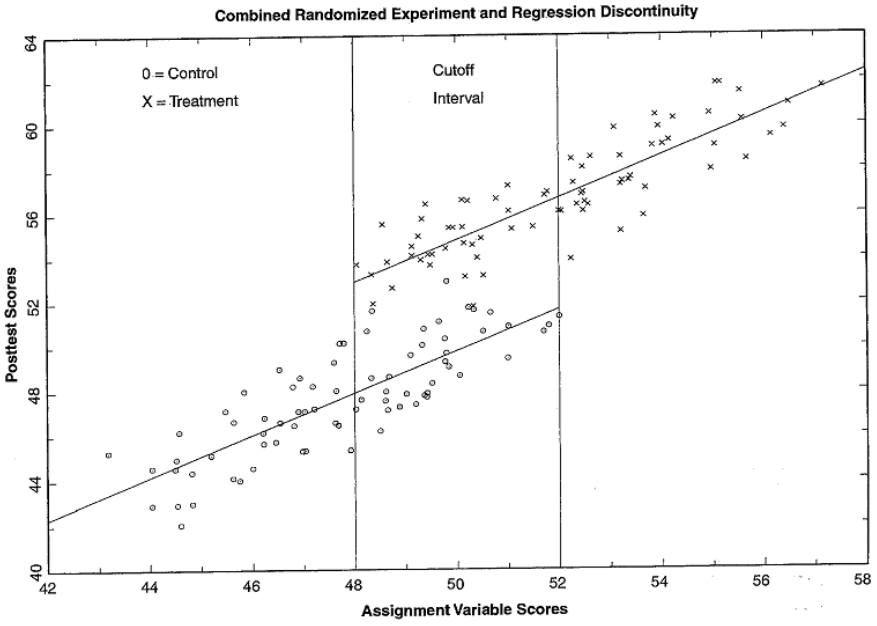
[424] Selection – instrumentation

دقیق‌تر برآورد کنیم. در شکل ۷.۱۳، این کار معادل با کم کردن ۴۸ است، زیرا همه شرکت‌کنندگانی که مداخله را دریافت کردند، در یک طرف ۴۸ بودند. با این وجود، اگر برهم‌کنشی وجود نداشته باشد، استفاده از میانه بازه به عنوان نقطه براورد، توان آزمایش را افزایش می‌دهد (Trochim & Cappelleri, 1992). اگر فرضیه اولیه محقق درباره این برهم‌کنش غلط باشد، می‌توان تحلیل جایگزین دیگری انجام داد، مادامیکه آن را به روشنی به مثابه [تحلیلهای] اکتشافی درنظر بگیریم.

آزمایش تصادفی را می‌توان به طرق مختلفی با طرح RD ترکیب کرد (Boruch, 1975؛ Cappelleri, 1991؛ Trochim, 1984, 1990؛ Trochim & Cappelleri, 1992). اول، همه بیمارانی که در یک طرف برش قرار دارند را می‌توان نیازمند در نظر گرفته و درمان را روی آنها اعمال کرد، و بقیه شرکت‌کنندگان به صورت تصادفی به شرایط مختلف تخصیص داده شوند. دوم، می‌توان یک طرح RD را به یک آزمایش تصادفی که از نقطه برش کمّی‌شده صلاحیت استفاده می‌کند (مثلاً Cappellori & Trochim, 1994)، افزود. به جای اینکه افرادی که نمره برش صلاحیت موردنظر را کسب نکرده‌اند را از مطالعه خارج کنیم، می‌توان آنها را به عنوان شرایط کنترل ناپیوستگی رگرسیون حفظ کنیم. این کار می‌تواند باعث افزایش هزینه اندازه‌گیری شود، اما در عین حال می‌تواند باعث شود تا بدون نیاز به افزایش تعداد نفرات دریافت‌کننده درمان، توان را افزایش دهیم. سوم اینکه، می‌توان از بازه‌های برش متعددی که در طول متغیر تخصیص قرار گرفته‌اند، استفاده کرد. به عنوان مثال، یک بازه در میانه، یک بازه در انتهای بالایی، و یک بازه در انتهای پایینی انتخاب کرد. این کار دو مزیته خواهد داشت: (۱) افزایش تعمیم‌پذیری اثر در طول طیف متغیر تخصیص- و نه تنها در محل نقطه برش؛ و (۲) استفاده از تصادفی‌سازی در نقاطی روی خط رگرسیون، که در آنها تأثیر اثرات کف و سقف در براورد خط رگرسیون مسئله‌ساز شده. چهارم، اینکه می‌توان شرکت‌کنندگان را به صورت تصادفی به بازه‌های متعددی تخصیص داد، به گونه‌ای که احتمال دریافت مداخله در طول بازه‌های مختلف متفاوت باشد. به عنوان مثال، سه بازه برش با احتمال ۲۵:۷۵، ۵۰:۵۰ و ۷۵:۲۵ تخصیص یافتن به گروه مداخله، میان احتمالهای تخصیص ۰:۱۰۰ و ۱۰۰:۰ (این احتمالات، بخش ناپیوستگی رگرسیون طرح هستند)، قرار می‌گیرند. این تغییرات، نگرانیهای مرتبط با عدم‌تخصیص صحیح افراد با نمرات مشابه را کاهش می‌دهد، زیرا هر چه نمره نیاز یا صلاحیت شرکت‌کننده (بر اساس متغیر تخصیص) بیشتر باشد، شانس او برای دریافت مداخله بالاتر خواهد بود. پنجم اینکه، احتمال تخصیص تصادفی درون بازه، می‌تواند در طول جایگاه‌های مختلف یا در درون جایگاه‌ها تغییر کند، و همینطور اندازه بازه‌ها ممکن است در طول یا در درون جایگاه‌های مختلف در طول زمان تغییر کند. ششم اینکه، بازه تصادفی‌سازی را می‌توان در نقطه‌ای قرار داد که در آن احتمال انحنا وجود دارد، تا به این وسیله مشکلات ناشی از تابعی مدلسازی را به حداقل رساند. هفتم، می‌توان شرکت‌کنندگان موجود در بازه برش را به صورت تصادفی تخصیص داد، اما همه دیگر شرکت‌کنندگان را تنها به یک طرف تخصیص داد. مثلاً در مورد داروهای خطرناک، می‌توان به همه شرکت‌کنندگان دیگر دارونما داد تا خطر برای آنها به حداقل برسد. و

در حالت وجود درمانی که احتمالاً موثر است، می‌توان به همه افراد خارج از بازه دارو داد، تا حداکثر فایده به آنها رسانده شود. برای کسب جزئیات دقیقتر راجع به هر کدام از این طرحها، و توان و نحوه انجام تحلیلها در آنها، خوانندگان باید به مقالات اصلی مراجعه نمایند (Cappelleri, 1991؛ Cappelleri, Darlington and Trochim, 1994؛ Trochim, 1984, 1990 و Trochim and Cappelleri, 1992 مراجعه کند).

از جمله دیگر مزایای طرح ترکیبی آن است که تصادفی‌سازی، قدرت آزمون اثرات مداخله را افزایش می‌دهد. این افزایش وابسته به اندازه افزایشی بازه تصادفی‌سازی، نسبت به طیف کلی متغیر تخصیص است (Cappelleri, 1991؛ Trochim & Cappelleri, 1992).



شکل ۷.۱۳. ترکیب آزمایش تصادفی و ناپیوستگی رگرسیون.

دومین مزیت آن است که طرح ترکیبی امکان تخمین خط رگرسیون گروه مداخله و کنترل را در طول طیف نمرات تخصیص مشابه، در درون بازه تصادفی‌سازی، فراهم می‌نماید. این امکان مشکل طرح RD پایه که در آن باید خطوط را بر نمرات برش تصویر کرد[۴۲۵] را کاهش می دهد، تصویرسازی‌ای[۴۲۶] که نیازمند مدلسازی دقیق

[425] Project
[426] Projection

شکل تابعی رگرسیون است. هم‌پوشانی خطوط رگرسیون در بخش تصادفی طرح، به ارزیابی دقت این تصویرسازی کمک می‌کند.

مزیت سوم طرحهای ترکیبی در توانایی آنها برای بهبود مشکلات عملی موجود در حالتیست که مشخص نیست محل نقطه برش کجا باید باشد. در مطالعات پزشکی، معمولاً مشخص است که همه شرکت‌کنندگان پایینتر از نقطه‌ای معین، به مداخله نیازی ندارند، و افراد بالاتر از آن حد بالایی معین، نیازمند مداخله هستند؛ اما نیاز افراد قرار گرفته در میانه این بازه به مداخله معلوم نیست. تخصیص تصادفی می‌تواند اخلاقی‌ترین راه برای تخصیص شرکت‌کنندگان درون این بازه باشد. بالعکس، در بسیاری از آزمایش‌های تصادفی، آنهایی که بیشترین نیاز را دارند کنار گذاشته می‌شوند، زیرا این افراد باید قطعاً درمان را دریافت کنند، و آنهایی که نیاز کمتری دارند نیز معمولاً در ابتدای راه کنار گذاشته می‌شوند، و آزمایش تنها روی شرکت‌کنندگانی انجام می‌شود که میزان نیاز آنها در بازه محدودی قرار می‌گیرد. با استفاده از طرح ترکیبی که در آن نیازمندترین و بی‌نیازترین شرکت‌کنندگان نیز در مطالعه وارد می‌شوند (افرادی که در صورت عدم بکارگیری طرح ترکیبی حذف می‌شدند)، توان آماری کل آزمایش افزایش می‌یابد (Luft, 1990).

از جمله دیگر تبیینهای طرح RD، که باز در ارتباط با طرحهای ترکیبی قرار دارد، ضرورت انجام آزمایش تصادفی ناقض-تساوی[۴۲۷] است. در شکل ۷.۱، تنها شرکت‌کنندگانی را در نظر بگیرید که بسیار نزدیک به نمره برش ۵۰ قرار می‌گیرند، یعنی آنهایی که بین ۴۹/۵ و ۵۰/۵ قرار دارند. این افراد شرکت‌کنندگانی هستند که بیشترین چالش در مورد تخصیص آنها به یکی از دو شرایط طرح RD وجود دارد. به عنوان مثال، مجریان طرح ممکن است بر این باور باشند که کودکی شایستگی دریافت مداخله را دارد، اما نمره او ۴۹/۵ است، و این نمره سبب حذف وی می‌شود. می‌توان برای این شرکت‌کنندگان یک آزمایش تصادفی انجام داد. اگر میانگین گروه‌های کنترل و مداخله را تنها برای شرکت‌کنندگانی که در این بازه قرار می‌گیرند رسم کنیم، همان فاصله‌ای را میان میانگین‌ها مشاهده خواهیم کرد که در صورت بکارگیری طرح ناپیوستگی رگرسیون حاصل می‌شد. بنابراین RD را می‌توان به صورت طرحی که خطوط رگرسیون را به نتایج این آزمایش تصادفی ناقض-تساوی فرضی تصویر می‌کند در نظر گرفت، زمانی که بازه برش به صفر میل می کند.

ترکیب ناپیوستگی رگرسیونی و شبه آزمایش

در همه طرح‌های ترکیبی توصیف شده تا به اینجا، می‌توان به جای آزمایش تصادفی، از شبه‌آزمایشها بهره گرفت. در آزمایش‌های پزشکی، افرادی که نمراتشان درون بازه برش قرار دارد، به صلاحدید پزشک یا تیم مداخله، درمان دریافت می‌کنند، اما کسانی که بالا یا پایین بازه هستند، به یک طرح RD تخصیص داده می‌شوند. در آموزشهای

[427] Tie-breaking

جبرانی، مسئولین اجرایی تصمیم می‌گیرند که به کدامیک از کودکان موجود در بازه برش مداخله تخصیص پیدا کند. در روان‌درمانی، مشتری به شخصه می‌تواند انتخاب کند که درمان را دریافت کند یا نه.

این ترکیب شبه‌آزمایش با طرح ناپیوستگی رگرسیون به ویژه در زمانی مفید است که تصمیم بر آن شده که طرح شبه‌آزمایش برای گروه مقایسه غیرهم‌ارز اجرا شود. معمولاً افزودن طرح RD به این شبه‌آزمایش، برای این مجریان، پزشکان و روان‌درمانان-که مایل به داشتن اختیاراتی که در جریان شبه‌ازمایشها به خودشان و یا مشتریانشان داده می‌شود، هستند- امری مطلوب است. با این همه، حتی در شبه‌آزمایش مطلوب نیز این احتمال وجود دارد که معیارهایی برای استثناء کردن برخی افراد وجود داشته باشد. چرا برخی از شرکت‌کنندگان کنار گذاشته می‌شوند؟ هم توان آماری و هم برآوردهای بدست آمده از شبه‌آزمایشها، با افزودن مؤلفه RD به طرح مطالعه – که تنها هزینه آن کمّی کردن معیارهای حذف، و محاسبه متغیرهای خروجی و تخصیص برای آن دسته از شرکت‌کنندگانی است که در غیر این صورت به هر حال حذف می‌شدند- بهبود می‌یابد.

همچنین می‌توان دیگر مؤلفه‌های طرح شبه‌آزمایش را با طرح RD ترکیب کرد. می‌توان یک طرح RD سنتی را اجرا کرده، سپس در انتهای مطالعه RD، به شرکت‌کنندگان گروه کنترل، مداخله داد. یا می‌توان یک طرح ناپیوستگی رگرسیون را روی گروه کنترل غیرهم‌ارزی که پیش‌بینی می‌شود اثری نشان ندهند (چون مداخله دریافت نکرده‌اند) اجرا کرد (مثلاً Cullen et al., 1999). یا اگربواسطه یک تهدید روایی درونی جدّی، مداخله تحت بررسی از دیگر مداخلاتی که از همان نقطه برش استفاده می‌کنند متمایز باشد، می‌توان تحلیل را با داده‌های قبل از مداخله دوباره اناجم داد تا نشان داده شود که هیچ ناپیوستگی‌ای قبل از مداخله وجود نداشته است، و یک ناپیوستگی بعد آن ظاهر شده است (همانگونه که پیشتر در مورد مطالعه اثرات شایستگی مدیکید روی تعداد ویزیت‌های پزشکان توضیح داده شد).

ناپیوستگی رگرسیون – آزمایش یا شبه آزمایش

هم در این کتاب و هم در کارهای پیشین (Campbell & Stanley, 1966؛ Cook & Campbell, 1979؛ Cook & Shadish, 1994) ما طرح ناپیوستگی رگرسیونی را در قالب یک شبه‌آزمایش طبقه‌بندی کردیم. این امر به علّت ادراک نویسندگان این کتاب از شبه‌آزمایشها است (شبه آزمایش طرحی است که ویژگی‌های ساختاری یک آزمایش را دارد اما فاقد تخصیص تصادفی است). بر اساس این مفهوم‌پردازی، طرح RD یک شبه‌آزمایش است. به علاوه، دلایل موجّه دیگری هم وجود دارد که باعث شود نتوانیم RD را برابر با یک آزمایش تصادفی قلمداد کنیم. اول اینکه، این طرح جدیدتر است، و نقایص طراحی و تحلیلی موجود در آن هنوز به اندازه آزمایش‌های تصادفی روشن نیست (Ahn, 1983؛ Cappelleri, Trochim, Stanley & Reichardt, 1991؛ Reichardt, Trochim & Cappellari, 1995؛ Robbins & Zhang, 1988, 1989, 1990؛ J. Shapiro, 1984؛ T. Stanley, 1991؛ T. Stanley & Robinson, 1990؛ Trochim, Cappelleri & Reichardt, 1991؛ Visser & deLeeuw, 1984؛ Williams, 1990). دوم اینکه مدلسازی دقیق

شکل تابعی رابطه بین متغیر تخصیص و متغیرهای خروجی در طرحهای RD نقش حیاتی دارد، و این امر نیازمند بکارگیری تحلیل‌های آماری دشوارتر - در مقایسه با تحلیل های موردنیاز در آزمایش‌های تصادفی- است. سوم، اینکه طرح RD نسبت به آزمایشهای تصادفی توان کمتری دارد (Cappelleri et al., 1994)؛ این مسأله اساساً به علّت وجود هم‌خطّی[۴۲۸] متغیرهای تخصیص و مداخله است، به ویژه زمانی که مقدار برش در چارک سوم یا پایین‌تر قرار داشته باشد (Cappelleri, 1991). اما حتی وقتی برش‌ها در محل میانگین باشند نیز، یک طرح RD برای یک اثر کوچک، ۲/۷۳ برابر بیشتر از یک آزمایش تصادفی متوازن، شرکت‌کننده لازم دارد تا بتواند به توان ۸۰٪ برسد؛ برای اندازه اثرهای متوسط، این نسبت چیزی در حدود ۲/۵۴ برابر، و درمورد اندازه اثرهای بزرگ این نسبت تا ۲/۳۴ کاهش می‌یابد (Cappelleri et al., 1994). استراتژی‌های مختلف افزایش توان آماری در طرحهای RD در جدول ۷.۲ ارائه شده است. البته، توصیه‌های استاندارد در مورد افزایش توان، مانند افزایش پایایی اندازه‌گیری و اندازه اثر، در طرحهای RD نیز کاربرد دارد.

با این وجود، برخی از دیگر نویسندگان بر این باورند که طرحهای RD چیزی بیش از یک شبه‌آزمایش است. به عنوان مثال، موستلر (Mosteller, 1990) چنین بیان می‌کند که «طبق تعریف این نویسنده – که در آزمایش، محقق کاربرد مداخلهها را کنترل می‌کند – طرح ناپیوستگی رگرسیون، در واقع یک آزمایش است» (صفحه ۲۲۵). همچنین، اگر بتوان آزمایش را به عنوان هر طراحی تعریف کرد که می‌تواند برآوردی بدون‌سوگیری از اثرات مداخله دهد، آنگاه طرح RD بر اساس این تعریف، باز هم یک آزمایش است. ما در فصل بعد این کتاب، وقتی دوباره به مفهوم آزمایش تصادفی برمی‌گردیم، این شباهت‌ها را دقیق‌تر بیان می‌کنیم.

جدول ۷.۲. روشهایی برای افزایش توان در طرحهای ناپیوستگی رگرسیونی

۱. مقدار برش را در محل میانگین متغیر تخصیص قرار دهید.
۲. از یک متغیر تخصیص پیوسته استفاده کنید، و یا از متغیری بهره بگیرید که سطوح ترتیبی بیشتری دارد.
۳. متغیر تخصیصی انتخاب کنید که کمترین همبستگی ممکن را با مداخله داشته باشد (گاهی می‌توان اینکار را با ایجاد ترکیبی از متغیرهای تخصیص مختلف انجام داد).
۴. مراقبت باشید که بیش از اندازه جملات برهم‌کنشی و غیرخطی را به مدل تحلیلی برازش نکنید؛ افزودن جملات غیرضروری تنها باعث کاهش درجه آزادی خواهد شد.
۵. آنهایی که در وسط بازه برش قرار می‌گیرند را به یکی از شرایط تخصیص دهید، و آنهایی که در دو طرف دیگر بازه قرار می‌گیرند را به شرایط دیگر؛ اینکار مسأله هم‌خطی بین متغیر تخصیص و متغیر مداخله را کاهش می‌دهد.

[428] Collinearity

۶. RD را با یک آزمایش تصادفی ترکیب کنید؛ زیرا هر چه مطالعه از توان بیشتری برخوردار باشد، محدوده بازه برشی که شرکت‌کنندگان به طور تصادفی به آن منسوب می‌شوند، وسیعتر خواهد بود.

ضمیمه۷.۱. منطق اثبات‌های آماری مرتبط با ناپیوستگی رگرسیونی

نویسندگان مختلف اثبات‌هایی را ارائه کرده‌اند که نشان می‌دهد RD، برآوردهایی بدون‌سوگیری از آثار مداخله ارائه می‌دهد (Cappelleri, 1991؛ Goldberger, 1972a, 1972b؛ Rubin, 1977). جزئیات این اثبات‌ها از حوصله این کتاب خارج است، اما برای روشنتر شدن تفاوت میان طرحهای RD و دیگر انواع شبه-آزمایش‌ها بهتر است قدری به شرح منطق کلی این اثباتها بپردازیم. برخی از مباحث در ارتباط با نکات بیان شده درخصوص خطای اندازه‌گیری که در شکل ۷.۸ - به صورت گرافیکی برای موارد خاص نشان داده شده است- قرار دارند؛ و این می‌تواند یک نقطه شروع خوب باشد. ما با این فرض ساده‌سازی شروع می‌کنیم که تنها مداخله و متغیر تخصیص، می‌توانند بر خروجی تأثیر بگذارند (بعداً بحث را اندکی تعمیم خواهیم داد). اثبات‌های آماری تلاش می‌کنند نشان‌دهند که «ضریب رگرسیون جزئی جمعیت بدست آمده از رگرسیون میان خطای اندازه‌گیری (u) و متغیر مداخله که به طور بی‌نقص اندازه-گیری شده (Z)، با فرض کنترل کردن یا ثابت نگه داشتن متغیرهای کمکی (با دقت اندک) اندازه‌گیری شده (X)، برابر با صفر است: $B_{u,Z|X} = 0$ (Cappelleri, Trochim, Stanley & Reichardt, 1991؛ صفحه ۴۰۶؛ همچنین Cappelleri, 1991 صفحه ۱۷۸ را ببینید). این، از نظر مفهومی، بدین معناست که در هر نمره معین روی متغیر تخصیص (X)، تخصیص به صورت کامل و بی نقص شناخته شده است، بنابراین همبستگی جزئی بین خطای اندازه‌گیری برای متغیرهای تخصیص (u) و متغیر فرضی که نماینده تخصیص مداخله است (Z)، صفر می‌باشد (وقتی که X را بدانیم، Z را نیز می‌دانیم، و اطلاعات u چیزی بیش از Z در اختیار ما نمی‌گذارد (در اینجا از بحث Cappelleri, 1991 تبعیت کرده، و خطا را به جای e با u نشان می‌دهیم). زمانی که شرط دوم برقرار باشد (که در ناپیوستگی رگرسیون اینگونه است)، ضریب رگرسیون Z برآوردی بدون‌سوگیری از اثر مداخله خواهد بود. توجه کنید که این شرط برای آزمایش تصادفی برقرار است، زیرا در هر سطح معینی از پیش‌آزمون (X)، تخصیص به مداخله (Z) به صورت تصادفی تعیین می‌شود و بنابراین، Z تنها به صورت تصادفی با متغیرهای دیگر از جمله دو مولفه X (نمره واقعی T و خطای اندازه‌گیری u) ارتباط دارد.

بالعکس، در داده‌های همبستگی و در طرح‌های گروه کنترل غیرهم‌ارز، حداقل برای سطوحی از X، اغلب خطای اندازه‌گیری متغیرهای کمکی با تخصیص به مداخله همبستگی دارد. این مسئله در شکل ۷.۸ نشان داده شده است. فرض کنید که مکانیسم انتخاب صحیح، به جای نمره مشاهده شده X بر اساس نمره مجهول و مشاهده نشده صحیح T باشد. به عنوان مثال، هر شرکت‌کننده‌ای که T او در پیش‌آزمون، ۵۶ یا بیشتر است، برای گروه مداخله انتخاب می‌شود. اگر X پیش‌آزمون خطایی نداشته باشد (یعنی مقیاس کاملی از T باشد)، تنها با نگاه کردن به

نمرات پیش‌آزمون می‌توانیم دریابیم که شرکت‌کننده به چه شرایطی تخصیص پیدا کرده است؛ این حالت در شکل‌های a۷.۸ و c۷.۸ دیده می‌شود. اما در طرح گروه کنترل غیرهم‌ارز، نمی‌توانیم T را مشاهده کنیم؛ تنها می‌توانیم متغیر کمی X که به شکل ضعیفی اندازه‌گیری شده است را به عنوان مقدار معرف برای مکانیسم انتخاب واقعی T در نظر بگیریم. برخی از شرکت‌کنندگان که به گروه مداخله تعلق دارند، چون نمرات واقعی آنها در شکل a۷.۸ از ۵۶ بیشتر است، با پیش‌بینی نادرست در گروه کنترل قرار خواهند گرفت، زیرا نمرات پیش‌آزمون مشاهده شده آنها در شکل b۷.۸ کمتر از ۵۶ بوده است. این تخصیص نادرست به این علت رخ می‌دهد که خطاهای منفی به برخی مقادیر واقعی بالای ۵۶ افزوده شده است، و موجب شده تا نمرات مشاهده شده کمتر از ۵۶ بشوند. به عبارت روشنتر، اگر تخصیص به گروه مداخله، بر اساس نمره صحیح T بیش از ۵۶ صورت باشد، و ما بگوییم که تنها شرکت‌کنندگانی که نمرات مشاهده شده (X) آنها بیشتر از ۵۶ است در گروه مداخله قرار می‌گیرند، مرتکب خطا شده‌ایم، و مشخصاً این اشتباه ناشی از خطای اندازه‌گیری تصادفی بوده، و با آن همبستگی دارد. ضریب $B_{u,Z|X}$ به سادگی نشان می‌دهد که خطای ما تا چه اندازه بزرگ است.

حال این بحث را تعمیم می‌دهیم. در پاراگراف‌های پیش، فرض شد که u تنها متشکل از خطای اندازه‌گیری تصادفی است. در واقع، همان اثبات را می‌توان به حالتی که u جمله‌ای اختلال‌زا است که نه تنها خطای اندازه‌گیری، بلکه هر متغیر حذف‌شده دیگری که بر خروجی تأثیر دارد را دربرمی‌گیرد، نیز تعمیم داد. این مبحث عمومی‌تر را می‌توان به صورت زیر بیان کرد (مطالعه Mohr, 1988, 1995، جزئیات بیشتری را ارائه کرده است). برخی از متغیرهای حذف‌شده، با متغیر خروجی رابطه دارند، اما با متغیرهای پیش‌بین معادله رگرسیون همبستگی ندارند. این متغیرها مشکلی در برآورد اثر مداخله ایجاد نمی‌کنند، زیرا مکانیسمی برای تأثیرگذاری بر ضرایب رگرسیون مرتبط ندارند؛ این حالت کاملاً مشابه آزمایش تصادفی است. برخی دیگر متغیرهای حذف‌شده هم با متغیرهای خروجی همبستگی دارند، و هم با برخی متغیرهای معادله رگرسیون. اگر نمره آزمون پیشرفت تحصیلی یک کودک به عنوان متغیر تخصیص در نظر گرفته شده باشد، تحصیلات مادر وی به روشنی با متغیر خروجی همبستگی داشته، و بر آن تاثیر می‌گذارد. اما برای اینکه دریابید که چرا این مسأله در RD مشکل‌ساز نمی‌شود، باید بین بخشی از متغیرهای حذف‌شده که با متغیر تخصیص همبستگی دارند، و آنهایی که همبستگی ندارد، تمایز قائل شویم. بر اساس تعریف، متغیرهای غیرهمبسته نمی‌توانند بر ضرایب رگرسیون متغیر تخصیص تأثیر بگذارند. تأثیر متغیرهای همبسته با متغیرهای خروجی نیز پیشتر، در ضریب رگرسیون متغیر تخصیص لحاظ شده است. این بدین معنی است که ضریب رگرسیون در اینجا، نشان دهنده اثرمخلوط متغیر تخصیص و متغیر حذف‌شده است؛ از این رو، براوردی فاقدسوگیری از اثر متغیر تخصیص روی نتایج نیست. اما این امر مشکل‌ساز نیست، زیرا ضریب رگرسیون متغیر تخصیص، ضریب مد نظر ما نیست؛ بلکه ضریب رگرسیون متغیر فرضی مداخله، موردنظر ماست. بنابراین، متغیر خذف‌شده در طرح RD، چه تاثیری بر ضریب رگرسیون مداخله خواهد داشت؟ هیچ تأثیری ندارد.

آن بخش از متغیرهای حذف‌شده که با X همبستگی دارد، پیشتر از ضریب Z خارج شده است؛ این امر در ضریب X نیز منعکس شده است. آن بخش از متغیر حذف شده که با X همبستگی ندارد، با Z نیز همبستگی نخواهد داشت، زیرا Z زمانی به طور کامل معلوم می‌شود که X معلوم باشد.[۴۲۹] توجه کنید که این بخش ناهمبسته از متغیر حذف‌شده، در جمله خطای معادله رگرسیون وارد شده است، بنابراین، جمله خطا با متغیر فرضی مداخله همبستگی ندارد (دقیقاً همانند آزمایش تصادفی).

لازم به ذکر است که نکته کلیدی منطق بیان شده این بود که Z را می توان به طور کامل با X پیش‌بینی کرد؛ اما این حالت برای زمانهایی که انتخاب به صورت کامل بر مبنای متغیر تخصیص صورت نمی‌گیرد (همانگونه که در طرح‌های گروه کنترل غیرهم‌ارز دیگر هم دیده می‌شود)، مصداق ندارد. در این موارد، بخش‌هایی از متغیرهای حذف‌شده در جمله خطا که با X همبستگی ندارند، می‌توانند همچنان با Z همبسته باشند (یا مخلوط شوند). از آنجا که رگرسیون OLS *همواره* ضرایب بتا را برای کمینه کردن این همبستگی انتخاب می‌کند، آن وزنها به میزانی که اینطور همبستگی‌ها در داده‌ها موجود باشد، نادرست خواهد بود. زیرا در این حالت، داده‌ها پیشفرضهای لازم برای مدل OLS را ندارند. در نتیجه وزن بتای حاصله برای اثر مداخله نیز واجد سوگیری خواهد بود.

[۴۲۹] توجه کنید که این گفته بدین معنی نیست که Z همبستگی کامل با X دارد؛ در واقع، بهتر است که تا زمانی که هر عامل باقیمانده‌ای به شانس بستگی دارد، Z همبستگی کامل با X نداشته باشد. در حالت حدی، وقتی که X و Z همبستگی صفر دارند، تخصیص مبتنی بر X کاملاً تصادفی است. به عنوان مثال، X می‌تواند یک پرتاب سکه یا پرتاب تاس باشد.

فصل ۸

آزمایشهای تصادفی: منطق، طرحها و شرایط لازم برای انجام آنها

تصادفی (رندم): [از کلمه شانسی، به صورت تصادفی با سرعتی بالا، از ریشه کلمه رندون[۴۳۰] به معنی سرعت و خشونت؛ و از ریشه کلمه رندیر[۴۳۱] در فرانسه قدیم به معنی دویدن] ۱. نداشتن یک الگو، هدف و یا برنامه خاص. ۲. در آمار. وجود شانس یا احتمال برابر یا یکسان رخ دادن برای هر یک از اعضای گروه

آیا می‌توان گفت که انجام مداخلات پیش‌دبستانی در مورد کودکان محروم زندگی آینده آنها را بهبود می‌بخشد؟ آزمایش برنامه پیش دبستانی پری که در سال ۱۹۶۲ انجام شد، به بررسی این سؤال می‌پرداخت. ۱۲۸ کودک آفریقایی- آمریکایی کم درآمد به طور تصادفی به دو موقعیت دریافت برنامه پیش-دبستانی و یا بدون-مداخله تخصیص داده شدند. نود و پنج درصد شرکت‌کنندگان تا سن ۲۷ سالگی دنبال شدند، و نتایج نشان داد که گروه مداخله به طور معناداری بهتر از گروه کنترل در زمینه اشتغال، فارغ‌التحصیلی از دبیرستان، سوابق دستگیری، دریافت کمک هزینه‌های اجتماعی، مالکیت خانه، و درآمدهای مالی عمل کرده بودند. اگرچه IQ و استعداد تحصیلی اولیه تا دوران جوانی حفظ نشده بود. این نتایج در کنار شواهد آزمایشی دیگر در زمینه اثرات مداخلات پیش-دبستانی، به فرماندار ایالتی کمک کرد تا حمایتهای سیاسی و مالی بیشتری را برای طرحهایی مانند

[430] Randon
[431] Randir

حمایتهای پیش-دبستانی[۴۳۲] تأمین نماید. در این فصل منطق زیربنایی، و انواع طرحهای آزمایشهای تصادفی مانند مثال بالا را ارائه خواهیم کرد. ضمناً شرایطی که تحت آن می‌توان اینگونه آزمایشها را با دشواری کمتری در فضای بیرون از آزمایشگاه انجام داد، را مورد تحلیل قرار می‌دهیم.

در علوم طبیعی، دانشمندان مداخلاتی را تحت شرایط عدم‌وجود هرگونه متغیری که مخدوش‌کننده[۴۳۳] اثرات مداخله باشند، به کار می‌گیرند. پس از آن نحوه تغییرات را مورد مشاهده قرار می‌دهند؛ برای مثال، اینکه آیا افزیش در میزان حرارات فشار گاز را افزایش می‌دهد. برای مطالعه این موضوع، گاز در محفظه‌ای قرار داده می‌شود تا از هر گونه عامل دیگری که می‌تواند بر فشار داخل اثر بگذارد دور نگه داشته شود. اما حتی در این مثال ساده نیز مداخله همچنان یک بسته مداخله ملکولی است که به دشواری می‌توان آن را واگشوده و توضیح داد. محفظه موردنظر از مواد مشخصی ساخته شده، گرما از نوع خاصی گرم‌کننده ساتع می‌شود، سطح معینی از رطوبت وجود دارد، و الی آخر. کنترل کامل و ایزولاسیون کامل و تام مداخله موردنظر محقق دشوار است، حتی در علوم طبیعی.

در بسیاری از تحقیقات اجتماعی، مشکلات کنترل بزرگتر و پیچیده‌تری وجود دارند که انجام موفقیت‌آمیز آزمایش را دشوار می‌سازند. برای مثال، غیرممکن است بتوان فردی را از خانواده جدا کرد تا بتوان اثر خانواده را حذف نمود. حتی در آزمونهای کشاورزی بر روی بذرهای جدید نیز، محدوده‌هایی که این بذرها در آنها کشت می‌شود را نمی‌توان از خاک و آبیاری جدا کرد. بسیاری از دانشمندان بر رویکردی از کنترل آزمایشی تکیه می‌کنند که متفاوت از ایزولاسیون فیزیکیست. این رویکرد تخصیص تصادفی نام دارد. ریشه تخصیص تصادفی به تحقیقات آماردانی به نام فیشر، که در زمینه کشاورزی تحقیق می‌کرد، برمی‌گردد (۱۹۲۵ و ۱۹۲۶ برای مطالعه تاریخچه این موضوع به مقاله Cowles, 1989 مراجعه نمایید). رندم‌سازی یا تصادفی‌سازی، پیش از این نیز مورد استفاده قرار گرفته بود (Dehue, 2000; Gosnell, 1927; Hacking, 1988; Hrobjartsson, Gotzche, & Gluud, 1998; McCall, 1923; Peirce & Jastrow, 1884; Richet, 1884; Stigler, 1986). اما فیشر منطق آماری آن را تبیین کرده، و استنباط علّی را به تصادفی‌سازی فیزیکی واحدها به شرایط مختلف در یک آزمایش ربط می‌دهد (Fisher, 1999).

نظریه تخصیص تصادفی

تخصیص تصادفی موجّه بودن تبیینهای جایگزین برای اثر مشاهده‌شده را کاهش می‌دهد. از این نظر، تخصیص تصادفی مانند دیگر مؤلفه‌های طرح، مانند پیش‌آزمون، همتایان[۴۳۴] و یا متغیرهای وابسته غیرهم‌ارز عمل

[432] Head start
433 Confounding variable
[434] Cohort

می‌کند. اما تخصیص تصادفی بواسطه یک مشخصه بسیار ویژه از دیگر مؤلفه‌ها متمایز می‌شود. مؤلفه‌ای که تنها در طرحهای ناپیوستگی رگرسیون مشابه آن وجود دارد، و عبارتست از اینکه تخصیص تصادفی می‌تواند تخمینهای بدون‌سوگیری از میانگین اثر مداخله[۴۳۵] بدست دهد (Rosenbaum, 1995a).[۴۳۶] بعلاوه، تخصیص تصادفی این کار را با صحت و دقت بیشتری نسبت به طرحهای ناپیوستگی رگرسیونی، و در طیف وسیعتری از کاربردها انجام می‌دهد. از آنجاکه استنباطهای علّی کارا و بدون سوگیری هدف تحقیق آزمایشی است، ضروریست تا محققین تخصیص تصادفی را بخوبی درک نموده، و با نحوه کار آن آشنا شوند.

تخصیص تصادفی چیست؟

تخصیص تصادفی را می‌توان هر رویه‌ای دانست که طی آن، واحدها تنها براساس شانس به شرایط مختلف تخصیص داده می‌شوند، و هر واحد احتمالی غیرصفر برای تخصیص‌داده‌شدن به یک شرایط را دارد. یکی از رویه‌های تخصیص جاافتاده پرتاب تاس است. در پرتاب هر سکه متوازنی، ۵۰٪ احتمال این وجود دارد که رو بیاید. در آزمایشی با دو شرایط آزمایشی، اگر برای واحدی رو بیاید، آن واحد به شرایط مداخله تخصیص داده می‌شود، و اگر پشت بیاید به شرایط کنترل. رویه‌ی دیگر پرتاب تاس است که شش وجه دارد. هر وجه تاس یک ششم احتمال آمدن دارد، اما اینکه دقیقاً کدام شماره بیاید کاملاً وابسته به شانس است. در ادامه این فصل رویه‌های رسمی‌تری مانند استفاده از جدول اعداد تصادفی را معرفی خواهیم کرد. اما پرتاب سکه و تاس مقدمه خوبی برای ورود به بحث تصادفی‌سازی هستند.

تخصیص تصادفی هم معنی نمونه‌گیری تصادفی نیست. ما در نظرسنجی‌های عمومی، هنگامی که از اعضای نمونه تصادفی درباره نظراتشان سؤال می‌کنیم، به طور شانسی نمونه‌های تصادفی از جمعیت می‌گیریم. نمونه‌گیری تصادفی تضمین می‌کند که پاسخهای بدست‌آمده از نمونه نزدیک به آن‌چیزی باشد که در صورت پرسش از تمامی اعضای جامعه بدست می‌آمد. تخصیص تصادفی در مقابل استنباط علّی را از طریق شبیه‌کردن

---

[435] Average treatment effect

[۴۳۶] سه مشاهده در مورد عبارت «تخمینهای بدون سوگیری از اثرات مداخله متوسط» قابل بحث است. اول، برخی آماردانها ترجیح می دهند مزیت تصادفی سازی را بدست آوردن یک تخمین زننده ثابت (چیزی که با افزایش اندازه نمونه به پارامتر جامعه ی خود منطبق شده یا می گراید) تعریف کنند؛ علی الخصوص به این دلیل که ما هیچگاه تعداد بی نهایت نمونه بر اساس نظریه انتظارات (که در همین فصل توضیح داده خواهد شد) نداریم. در اینجا کلمه «بودن سوگیری» اولاً به این دلیل مورد استفاده قرار گرفته که برای خوانندگان غیرآماری قابل درکتر است چون به شکل بهتری با منطق کیفی کنترل سوگیری که زیربنای نوع شانسی ما در مورد روایی را تشکیل می دهد، تناسب دارد. دوم آنکه در یک مدل تخصیص تصادفی، میانگین نمونه همواره تخمنی بدون سوگیری از میانگین جمعیت است، بنابراین تفاوتهای میان میانگینهای نمونه همواره تخمینهای بدون سوگیری از تفاوتهای میان میانگینهای جمعیت هستند. تخمین از میانگین جمعیت را می توان بدون تخصیص تصادفی نیز بدست آورد. اما اینگونه تخمینها همانند تخمینهای بدون سوگیری از اثرات مداخله نیستند. و این دقیقاً کاری است با تخصیص تصادفی تسهیل می شود. سوم، این عبارت به میانگین اثر بر واحدهای حاضر در یک مطالعه برمی گردد، به گونه ای متمایز از اثرات بر هر کدام از واحدهای در مطالعه، که در آزمایش تصادفی مورد آزمون قرار نمی گیرد.

نمونه‌ها به یکدیگر انجام می‌دهد؛ در حالی که نمونه‌گیری تصادفی یک نمونه را شبیه به یک جامعه می‌نماید. این دو رویه در مفهوم تصادفی بودن مشترک هستند، اما هدف از این تصادفی‌بودن کاملاً متفاوت است.

چرا تصادفی سازی سودمند است؟

ادبیات موجود تبیین‌های آماری و مفهومی متعددی را برای چرایی و چگونگی تسهیل‌کننده‌بودن استفاده از تخصیص تصادفی در بدست‌آوردن استنباطهای علّی ارایه می‌کند. از آن جمله می‌توان به موارد زیر اشاره داشت:

- تخصیص تصادفی تضمین می‌کند که تبیین‌های جایگزین علّی مخدوش‌کننده اثر مداخله اعمال‌شده بر روی واحدهای حاضر در یک شرایط نبوده‌اند.
- تخصیص تصادفی موجّه‌بودن تهدیدهای روایی را با توزیع تصادفی آنها میان شرایط مختلف کاهش می‌دهد.
- تخصیص تصادفی گروهها را از نظر مقدار موردانتظار تمامی متغیرها در پیش‌آزمون (خواه محاسبه‌شده باشند و خواه نه) یکسان می‌سازد.
- تخصیص تصادفی به محقق اجازه می‌دهد تا فرایند انتخاب را به درستی شناخته و مدل نماید.
- تخصیص تصادفی امکان محاسبه تخمین معتبری از واریانس خطا را برقرار می‌نماید؛ تخمینی که همچنین متعامد[۴۳۷] با مداخله باشد.

این تبیین‌های به ظاهر متفاوت در واقع کاملاً به یکدیگر مرتبط‌اند. هیچکدام از آنها به خودی خود نمی‌تواند نمای کاملی از آنچه که تخصیص تصادفی انجام می‌دهد ارائه نماید، اما هر کدام بخشی از آن را انعکاس می‌دهند.

تخصیص تصادفی و تهدیدات روایی درونی

اگر گروههای مداخله را بتوان پیش از مداخله مشابه کرد، و آنها بعد از اعمال مداخله متفاوت بشوند، می‌توان گفت که تفاوتهای ناشی از انتخاب پیش‌آزمون، علّت تفاوتهای مشاهده‌شده در پس‌آزمون نبوده اند. با در نظر گرفتن گروههای مساوی در پیش‌آزمون، گروه کنترل پس‌آزمون را می‌توان به عنوان منبعی برای استنباط نقیض[۴۳۸] برای گروه مداخله پس‌آزمون در نظر گرفت (البته در محدوده‌ای که بعداً به آن اشاره خواهیم داشت). در نظر داشته باشید که منطق استنباط علّی اینجا در جریان است. ساختار زمانی آزمایش تضمین‌کننده آن است که علّت قبل از اثر رخ داده است. اینکه آیا علّت با اثر کوواریانس دارد، را می‌توان به راحتی در داده‌ها بررسی

[437] Orthogonal
[438] Counterfactual

نمود. تنها کار باقی مانده این است که نشان دهیم، اغلب تبیین‌های جایگزین برای رابطه علّت-اثر مشاهده‌شده، غیرموجّه هستند. تخصیص تصادفی اینکار را بوسیله توزیع تصادفی این تهدیدها میان شرایط مختلف، انجام می‌دهد. در نتیجه واحدهای موجود در شرایط مداخله احتمالاً همان متوسط مشخصاتی را خواهند داشت که واحدهای موجود در شرایط کنترل. تنها تفاوت سیستماتیک میان این شرایط، مداخله خواهد بود.

برای مثال، مطالعه‌ای را در نظر بگیرید که اثرات رواندرمانی را بر استرس مورد بررسی قرار می‌دهد. استرس می‌تواند دلایل متنوعی مانند بیماری، تضادهای مادی، از دست دادن کار، جدل و مناقشه با همکاران، و مرگ والدین داشته باشد. حتی رویدادهای مثبت مانند بدست آوردن کار جدید، و یا ازدواج کردن می‌تواند استرس‌زا باشد. آزمایش‌کننده باید بتواند تضمین کند که هیچکدام از این علّتهای جایگزین، مخدوشگر اثر دریافت رواندرمانی نبوده است؛ زیرا در غیر این صورت، نمی‌توان گفت که آیا اثر مشاهده‌شده در پس‌آزمون به دلیل دریافت روندرمانی رخ داده یا یکی از این مخدوشگران. تخصیص تصادفی تضمین می‌کند که هر مراجع دریافت‌کننده رواندرمانی، می‌توانسته به اندازه هر یک از اعضای گروه کنترل، تحت‌تأثیر شغل جدید، و یا طلاق اخیر بوده باشد. تخصیص تصادفی نه از وقوع این دلایل جایگزین (مثلاً طلاق) پیشگیری یا جلوگیری می‌کند، و نه واحدها را نسبت به این رویداد ها ایزوله می‌کند. افراد در یک آزمایش تصادفی همچنان طلاق می‌گیرند و کار جدید پیدا می‌کنند؛ تخصیص تصادفی تنها تضمین می‌کند که اینگونه رویدادها با احتمال بیشتری برای افراد حاضر در گروه مداخله (در مقایسه با گروه کنترل) رخ نمی‌دهند. در نتیجه، اگر در پس‌آزمون، مراجعین رواندرمانی در مقایسه با گروه کنترل، استرس کمتری را گزارش کنند، کمتر احتمال دارد که علّت تفاوت مشاهده‌شده این باشد که یک گروه موقعیتهای جدید شغلی یا طلاق بیشتری داشته؛ چون احتمال وقوع این عوامل استرس‌آور در هر دو گروه برابر بوده است. تنها دلیل سیستماتیک توضیح‌دهنده تفاوت که می‌تواند نتایج را توضیح دهد، مداخله خواهد بود.

تخصیص تصادفی تنها می‌تواند از بروز یک تهدید روایی درونی جلوگیری کند، و آن سوگیری انتخاب است. اینکار بنا به تعریف تخصیص تصادفی انجام می‌شود، زیرا سوگیری انتخاب عبارتست از بکارگیری روشی واجد سوگیری سیستماتیک برای انتخاب واحدهای هر یک از شرایط، درصورتی که شانس (به عنوان عامل تخصیص) چنین سوگیری سیستماتیکی ندارد. تخصیص تصادفی نمی‌تواند از دیگر تهدیدات روایی، مانند بلوغ یا رگرسیون به میانگین واحدها جلوگیری کند؛ همچنین نمی‌تواند پس از شروع مطالعه از هیچ رویدادی غیر از مداخله پیشگیری کند (منظور تهدید گذشت زمان است). پیش‌آزمون همچنان می‌تواند نوعی اثر (تجربه)آزمون داشته باشد، و تغییر در ابزار همچنان می‌تواند رخ دهد. تخصیص تصادفی به سادگی احتمال آنکه این تهدیدها بتوانند با اثر مداخله مخلوط شوند را کاهش می‌دهد.

معادل کردن گروهها بر اساس انتظارات

در آمار، عموماً به گفتن این جمله که تخصیص تصادفی در پیش‌آزمون، گروهها را از نظر عناصر مورد انتظار مشابه کرده است، بسنده می‌شود. این جمله به چه معناست؟ اول، به این معنا نیست که تخصیص تصادفی واخدها را از نظر نمرات مشاهده‌شده پیش آزمون معادل می‌کند. هووارد و همکارانش (۱۹۸۶) به ما یادآوری می‌کنند که زمانی که یک دست کارت بازی ۵۲ تایی به خوبی مخلوط می‌شوند (بُر می‌خورند)، همچنان برخی بازیکنان دست بهتری نسبت به سایرین دریافت می‌کنند. این پدیده در میان بازیکنان به شانس دَست شهرت دارد (آماردانها آن را خطای نمونه‌گیری می‌نامند). در بازی ورق، ما انتظار نداریم که هر یک از بازیکنان در هر دست، کارتهایی عیناً به خوبی دیگران دریافت کند، اما انتظار داریم که کارتها در طول بلندمدت در میان دُور-دستهای متعدد (از نظر خوبی کارتها) مشابه باشند. تمامی این موارد در مورد آزمایش تصادفی نیز صدق می‌کند. در هر آزمایش نیز، زمانی که در برخی از شرایط آزمایشی، شرکت‌کنندگان بهتری وجود دارد، میانگینهای مشاهده شده پیش‌آزمون به دلیل شانس دست متفاوت هستند. اما ما می‌توانیم انتظار داشته باشیم که شرکت‌کنندگان در طول شرایط مختلف، و در جریان آزمایشهای تصادفی متعدد، طی بلندمدت، مساوی خواهند بود.

پس، به طور تکنیکی، تخصیص تصادفی گروهها را از نظر انتظارات نسبت به میانگین گروهها در پیش‌آزمون برابر می‌کند. به این معنی که بر اساس میانگین توزیع میانگین نمونه‌های ممکن بدست‌آمده از تمام تخصیص‌های تصادفی ممکن از واحدهای شرایط مختلف، برابر می‌کند. تصور کنید که محققی در یک مطالعه واحدها را به طور تصادفی به شرایط مداخله و کنترل تخصیص می‌دهد، و سپس میانگین تعدادی از متغیرهای نمونه را برای هر یک از شرایط محاسبه می‌کند. این دو میانگین تقریباً به طور قطع به دلیل خطای نمونه‌گیری (یا همان شانس دست) با یکدیگر متفاوت خواهند بود. اما فرض کنید که محقق اینکار را برای بار دوم تکرار کند، نتایج را یادداشت کرده، و این کار را برای دفعات بسیار زیادی انجام دهد. در پایان، محقق توزیعی از میانگینها برای گروه مداخله، و توزیعی از میانگینها برای گروه کنترل بدست خواهد آورد. برخی از میانگینهای گروه مداخله بزرگتر از دیگران هستند؛ همین امر برای گروه کنترل نیز مصداق دارد. اما متوسط تمامی میانگینها برای گروه مداخله، برابر خواهد بود با متوسط تمامی میانگینها برای گروه کنترل. بنابراین، کلمه انتظارات مستتر در تعریف تخصیص تصادفی، به میانگین تمامی میانگینها برمی‌گردد، و نه یک میانگین بدست‌آمده در مطالعه‌ای خاص.

زمانی که تفاوتهای تصادفی در میانگینهای مشاهده‌شده پیش‌آزمون وجود دارند، این تفاوتها نتایج مطالعه را تحت‌تأثیر قرار می‌دهند. برای مثال، اگر علیرغم تخصیص تصادفی، مراجعین تخصیص داده‌شده به رواندرمانی در ابتدای برنامه، افسرده‌تر از افراد تخصیص داده شده به گروه کنترل باشند؛ و اگر رواندرمانی افسردگی را کاهش داده باشد، نمرات پس‌آزمون ممکن است به دلیل تفاوتهای گروهی در پیش‌آزمون، در مرحله پس‌آزمون در گروههای کنترل و مداخله برابر باشند. در نتیجه، تفاوتهای پس‌آزمون میان گروههای مداخله و کنترل می‌تواند

اینطور تفسیر شود که مداخله اثری نداشته است، در حالی که به واقع داشته، و اثر مداخله بواسطه خطای نمونه‌گیری در تخصیص تصادفی، پوشانده شده است. به طور کلی، نتایج هر آزمایش تصادفی، به واسطه وجود شانس اینگونه تفاوتهای پیش‌آزمونی، تا حدودی متفاوت از اثرات در جمعیت است. در نتیجه، خلاصه (جمع‌بندی) نتایج آزمایشهای تصادفی متعدد بر روی یک موضوع (مانند آنچه در متا-آنالیز رواندرمانی انجام می‌شود)- در قیاس با آزمایشهای منفرد- می‌تواند تخمینهای دقیقتر و درست‌تری از اثرات مداخله به دست دهد. با همه این احوال، و با علم به وجود خطای نمونه‌گیری، همچنان می‌توان ادعا کرد که تخمینهای بدست آمده بر اساس یک مطالعه منفرد [تصادفی] نیز بدون سوگیری است. بدون سوگیری به این معناست که هرگونه تفاوت میان اثرات مشاهده شده و اثرات در جمعیت، نتیجه شانس بوده؛ و به این معنا نیست که اثرات یک مطالعه منفرد عیناً مطابق اثرات «واقعی» موجود در جمعیت است.

در توضیحات قبلی، از میانگین پیش‌آزمون برای نشان دادن نحوه کارکرد تصادفی‌سازی استفاده می‌شود. اگرچه، این بیشتر ابزاری برای تدریس است، و بکارگیری پیش‌آزمونهای واقعی ربطی به منطق [تصادفی‌سازی] ندارند. تصادفی‌سازی گروهها را از نظر انتظارات نسبت به هر متغیری (خواه مشاهده شده، و خواه نشده) قبل از مداخله مشابه می‌کند. در عمل البته، پیش‌آزمونها بسیار مفید هستند، زیرا امکان تشخیص بهتر و تعدیل برای ریزشها را فراهم می‌آورند؛ بکارگیری تکنیکهای آماری که توان آماری را افزایش می‌دهند را تسهیل می‌کنند؛ و می‌توان از آنها برای بررسی اینکه آیا مداخله در سطوح مختلف پیش‌آزمون به یک میزان اثربخش بوده یا نه، استفاده کرد.

دلایل آماری بیشتری برای اینکه چرا تخصیص تصادفی سودمند است

تصادفی‌سازی این را تضمین می‌کند که متغیرهای مخدوشگر، با مداخله دریافت‌شده توسط یک واحد (فرد)، همبستگی ندارند. یعنی اینکه سکه رو یا پشت بیاید، ربطی به اینکه شما مطلّقه، عصبی، پیر، مرد و یا هر چیز دیگری باشید، ندارد. در نتیجه، می‌توانیم پیش‌بینی کنیم که همبستگی پیش‌آزمونی میان تخصیص مداخله و متغیرهای مخدوشگر بالقوه نباید به طور معناداری متفاوت از صفر باشد.

این همبستگی صفر از نظر آماری کاربرد بسیاری دارد. برای درک این موضوع لازم است تا گریزی بزنیم به این بحث که چطور باید اثر مداخله را در مدلهای خطی تخمین زد. اول اجازه بدهید میان مطالعه و تحلیل مطالعه تمایز قائل شویم. در یک مطالعه بر روی اثرات رواندرمانی، استرس متغیر وابسته است (y_i)، رواندرمانی متغیر مستقل (z_i)، و مخدوشگران بالقوه در مقدار خطا (e_i) لحاظ شده و مستترند. در تحلیل این مطالعه، اثرات مداخله با استفاده از مدل خطی زیر تخمین زده می‌شوند.

$$y_i = \mu + \hat{\beta} z_i + e_i \qquad (۸.۱)$$

در این معادله μ ثابت معادله، $\hat{\beta}$ ضریب رگرسیونی، و اندیس i دامنه‌ای دارد از ۱ تا n (در جایی که n تعداد واحدهای حاضر در مطالعه هستند). بنابراین، y_i نمره‌ایست که واحد i ام از مقیاس استرس کسب کرده است، z_i برابر خواهد بود با ۱، اگر واحد موردنظر تحت رواندرمانی بوده، و صفر، اگر نبوده باشد. و در نهایت، e_i مشتمل است بر تمامی متغیرهای مخدوشگر بالقوه. در تحلیل، اگر $\hat{\beta}$ به طور معناداری متفاوت از صفر باشد، می‌توان گفت که روندرمانی اثر معناداری بر استرس داشته، و $\hat{\beta}$ بزرگی و جهت آن اثر را محاسبه می‌کند.

اگرچه، برای اینکه تمامی این موارد به درستی عمل نمایند، مدلی که در تحلیل تعیین[۴۳۹] می‌شود، باید با واقعیت مطالعه مطابقت داشته باشد. عدم موفقیت در کسب این تطابق، خطای تعیین[۴۴۰] نامیده می‌شود. خطای تعیین عبارتست از تعیین نادرست مدل، که انتظار می‌رود موجب افزایش داده‌ها شود. به بیان دقیقتر، تکنیکهای آماری بکار گرفته‌شده برای تخمین مدلها، مانند معادله ۸.۱، مقادیر ثابت $\hat{\beta}$ را به گونه ای انتخاب می‌کنند که همبستگی‌های میان خطاهای منتجه، و متغیرهای پیشبین، برابر با صفر باشد. آمار اینکار را فارغ از آنکه همبستگی موردنظر در مطالعه، در حقیقت صفر بوده باشد، انجام می‌دهد. خوشبختانه، تخصیص تصادفی تضمین می‌کند که همبستگی موجود در مطالعه، به دلایلی که در بخش قبلی به آنها اشاره شد، برابر با صفر خواهد بود؛ در نتیجه، مطالعه با تحلیلها انطباق دارد. اگرچه، در مطالعات غیرتصادفی، متغیرهای مخدوشگر بسیاری احتمالاً با دریافت مداخله همبستگی دارند؛ اما برنامه کامپیوتری همچنان $\hat{\beta}$ را به گونه‌ای انتخاب می‌کند که گویی خطا همبستگی حداقلی با پیشبین‌های موجود در تحلیل داده‌ها دارد. اینکار باعث می‌شود تا میان مطالعه و تحلیلها عدم‌انطباق وجود داشته باشد، و در نتیجه، نتایج تحلیل، تخمینی نادرست از اثر مداخله باشد[۴۴۱].

راهی دیگر برای بیان منافع تصادفی‌سازی آن است که بگوییم تصادفی‌سازی یک تخمین معتبر و روا از واریانس خطا ارائه می‌دهد (Keppel, 1991; Kirk, 1982). دو علّت ممکن است موجب مجموع واریانسها در نتایج شده باشند (در اینکه افراد چه میزان از نظر سطح استرس با یکدیگر تفاوت دارند)؛ یکی شرایط مداخله (اینکه آیا فرد رواندرمانی دریافت‌کرده یا نه)، و دیگری تمامی دیگر علل استرس. تخصیص تصادفی به ما اجازه می‌دهد تا این دو منبع واریانس را تفکیک کنیم. واریانس خطا در قالب میزان واریانس میان واحدها، در درون هر یک از شرایط تخمین‌زده می‌شود. برای مثال، برای مراجعینی که همگی به شرایط رواندرمانی تخصیص داده شده‌اند، واریانس ناشی از اینکه آیا رواندرمانی دریافت کرده‌اند یا نه نمی‌تواند منشاء تفاوتهای مشاهده‌شده در سطوح مختلف استرس آنها باشد. چون تمامی آنها رواندرمانی دریافت کرده‌اند، و در این زمینه، واریانس وجود نداشته است. در نتیجه هرگونه واریانس در نتایج میان افرادی که به طور تصادفی به رواندرمانی تخصیص‌یافته‌اند، می‌بایست

439 Specified
440 Specification error
[۴۴۱] یک راه برای اندیشیدن به مدلهای سوگیری انتخاب در فصل ۵ این است که این مدلها تلاش می کنند تا مولفه خطا را به شیوه ای قابل قبول از نظر آماری عمود بر پیش‌بین‌ها نمایند. اما این کار دشوار است؛ در نتیجه غالباً با شکست مواجه می‌شوند. همچنین یک راه برای اندیشیدن به طرحهای ناپیوستگی رگرسیونی آن است که این طرحها می‌تواند همبستگی را به دلایلی که در پیوست فصل ۷ ذکر شد، به صفر برسانند.

بواسطه متغیرهای مخدوشگر ایجاد شده باشد. متوسط هر کدام از این خطاهای محاسبه‌شده از درون هر یک از شرایط به عنوان بهترین تخمین ما از خطا عمل خواهد کرد. این خطا مبنایی است که بر اساس آن بررسی می‌کنیم که آیا تفاوت میان شرایط مداخله، بیشتر از تفاوتهاییست که به طور معمول میان واحدهای موجود در هر شرایط (تفاوتهایی که محصول تمامی دیگر علل ایجاد کننده استرس است) رخ می‌دهد؟

خلاصه

تخصیص تصادفی از بسیاری جهات می‌تواند استنباط علّی را تسهیل نماید. از طریق معادل کردن گروهها قبل از شروع مداخله، از طریق غیرموجّه‌کردن تبیینهای جایگزین، با خلق مؤلفه‌های خطا در معادله، که غیرهمبسته با متغیرهای مداخله هستند، و در نهایت از طریق فراهم آوردن امکان انجام تخمینهای روا و معتبر از این خطاها. این دلایل همگی به یکدیگر مرتبط هستند. برای مثال، زمانی که گروهها قبل از شروع مداخله معادل یکدیگر می‌شوند، امکان کمتری برای بروز عوامل مرتبط با تبیین‌های جایگزین وجود خواهد داشت، و غیرهمبسته بودن خطاها، برای تخمین اندازه مؤلفه خطا[۴۴۲] ضروریست. اما تصادفی‌سازی تنها راه برای محقق کردن این اهداف نیست. برخی اوقات می توان تبیینهای جایگزین را از طریق دیگر ابزارهای منطقی نیز غیرموجّه نمود (مانند آن چیزی که در شبه‌آزمایشها رایج است). خطاهای غیرهمبسته نیز می‌توانند با دیگر اشکال تخصیصهای کنترل‌شده مانند طرحهای رگرسیون ناپیوسته تولید شوند. اما تصادفی‌سازی تنها طرحی است که تمامی این اهداف را یکجا تأمین کرده؛ و اینکار را با روایی بیشتر و مشخصات بهتری نسبت به هر طرح جایگزینی انجام می‌دهد.

تخصیص تصادفی و واحدهای تصادفی‌سازی

ما به طور مکرر کلمه «واحد» را برای هر آن چیزی یا کسی که به شرایط مختلف در یک آزمایش تخصیص داده می‌شود بکار بردیم. یک واحد عبارتست از «فرصتی برای اعمال یا ترک یک مداخله» (Rosenbaum, 1995a, p.17).

انواع واحدها

در بسیاری از آزمایشها، واحدهای تخصیص داده شده به شرایط آزمایش، افراد هستند؛ مانند مراجعین به رواندرمانی، بیماران در آزمایشهای سرطان، یا دانش‌آموزان در مطالعات آموزشی. اما واحدها می‌توانند دیگر موجودات نیز باشند (Boruch & Foley, 2000). فیشر (R.A.Fisher, 1952) قطعات زمین را به طور تصادفی به سطوح مختلف کودها یا گونه‌های مختلفی از بذرها تخصیص داد. در تحقیقات پزشکی و روانشناسی، حیوانات

[442] Error term

اغلب به طور تصادفی به شرایط مختلف آزمایش تخصیص داده می‌شوند. محققین در آزمایش مالیات بردرآمد منفی نیوجرسی (Rees, 1974)، خانواده‌ها را به طور تصادفی به شرایط آزمایش تخصیص دادند. گازنل (Gosnell, 1927) اینکار را در مورد همسایه‌ها انجام داد. ادینگتون (۱۹۸۷) طرحهایی با یک شرکت‌کننده را مورد بحث قرار می‌دهد که در آنها، دفعات مداخله به طور تصادفی تخصیص داده می‌شود. مدارس به طور رندم تخصیص داده می‌شوند (Cook at al. 1998; Cook, Hunt & Murphy, 2000). رندم‌سازی تنها در علوم اجتماعی کاربرد ندارد. ویلسون (Wilson, 1952) مطالعه‌ای را تشریح می‌کند که در آن، صفحات فولاد مورداستفاده پیش از شروع آزمایش مواد منفجره مختلف رندم‌سازی شدند، به گونه‌ای که واریانسها در استحکام صفحات، ارتباط معناداری با هیچکدام از مواد منفجره نداشته باشد. بی‌نهایت احتمال وجود دارد.

واحدهای سطح بالاتر[۴۴۳]

واحدهایی مانند خانواده، محل‌های کار، کلاسهای درس، گروههای رواندرمانی، بیمارستانها، مناطق همسایه، و یا گروهها، گروههای مجتمعی از افرادی مانند اعضای خانواده، کارکنان، دانش‌آموزان، مشتریان، همسایگان، بیماران، همسایگان و یا شهروندان هستند. مطالعه تأثیر مداخلهها بر چنین واحدهای سطح بالاتری، در آزمایشات معمول است؛ و ادبیات ویژه‌ای حول انجام آزمایش در واحدهای سطح بالاتر شکل گرفته است (Donner & Klar, 2000; Murray, 1998; Sorensen, Emmons, Hunt & Johnston, 1998; Gail, Mark, Carrol, Green, & Pee, 1996). برای مثال، در آزمایش پرداخت آینده‌نگرانه آژانس ملی سلامت خانه، ۱۴۲ آژانس سلامت منزل به اشکال مختلف پرداخت تخصیص داده شدند، و تأثیر آن بر نحوه استفاده از خدمات مورد بررسی قرار گرفت (Goldberg, 1997)؛ آزمایش پرداختهای انگیزشی خانه پرستاری سن دیگو، ۳۶ خانه پرستاری را به گزینه‌های مختلف اشکال بازپرداخت هزینه‌های درمان تخصیص داد (Jones & Meiners, 1986)؛ آزمایش اندازه کلاس تنسی، ۳۴۷ کلاس درس را به صورت تصادفی به دو شرایط تعداد زیاد و اندک دانش‌آموزان تخصیص داد (Finn & Achilles, 1990)؛ و کلی و همکارانس (Kelly et al., 1997) هشت شهر را به دو شرایط مختلف تخصیص دادند تا نوعی درمان پیشگری از ایدز را مورد مطالعه قرار دهند. الزامی ندارد واحد سطح بالاتر به طور طبیعی وجود داشته باشد؛ مانند یک محیط کار، و یا یک منطقه جغرافیایی. محقق می‌تواند واحد سطح بالاتر را منحصراً برای اهداف تحقیق بسازد؛ مانند کاری که در برنامه ترک سیگار انجام شد، و در آن مداخله ترک سیگار به صورت گروهی، و به گروههای کوچکی از افراد ارائه می‌شد، و شرکت‌کنندگان می‌توانستند از حمایت متقابل اجتماعی بهره مند شوند. الزامی هم ندارد که افراد واحد سطح بالا یکدیگر را بشناسند، و یا با یکدیگر تعامل داشته باشند. به عنوان نمونه، زمانی که اقدامات پزشکی به صورت تصادفی به شرایط مختلف تخصیص داده

[443] Higher order units

می‌شود، اقدامات پزشکی یک واحد سطح بالاتر به حساب می‌آید، حتی اگر عمده بیماران پزشکان هیچ گاه یکدیگر را ملاقات نکرده باشند. در نهایت، برخی اوقات یک مداخله ماهیتاً قابل محدود شدن به یک فرد خاص نیست. مثلاً زمانی یک کمپین امنیت رانندگی رادیومحور از طریق یک محدوده رادیویی پخش می‌شود، کل منطقه تحت پوشش آن کانال رادیویی مداخله موردنظر را دریافت می‌کنند، حتی اگر تنها تعداد معدودی از رانندگان در مطالعه وارد شده باشند (Reicken et al., 1974).

غالباً دلایل قابل‌قبول علمی و کاربردی برای اســتفاده از واحدهای مجتمعی وجود دارد. در آزمایش یک کارخانه، احتمالاً منطقی نیسـت که هر یک از کارگران را ایزوله کرده، و به هریک از آنها یک مداخله منحصـر به فرد ارائه کنیم؛ چون اینکار می‌تواند مقاومت کارگران، و یا انتشار مداخله به دیگر کارگران را در پی داشته باشد. به همین ترتیب، در ارزیابی اولیه برنامه کوچه کنجد، محققان کودکان حاضــر در مهدکودکهای مکزیکی را در قالب گروههای کوچک به تماشــای برنامه مذکور واداشـتند. این کودکان در اتاق مخصـوصـی بودند، و دو نفر مربی نظارت می‌کردند که کدامیک روی تماشـای برنامه تمرکز کرده است. به طور همزمان دیگر کودکان کارتونها را در اتاقهای بزرگتر در یک اتاق معمولی، و بدون هیچ پایش ویژه تماشا می‌کردند. از آنجا که رفتار غیرمشـابه با همکلاسـیها می‌تواند موجب یک نابرابری متمرکز شـده باشـد، بهتر می‌بود اگر منابع آزمایش‌کنندگان به آنها اجازه می‌داد تا کل کلاس را به مداخلهها تخصیص دهند.

سؤال تحقیق نیز می‌تواند تعیین‌کننده این باشد که چه سطحی از واحدها باید مورد تصادفی‌سازی قرار بگیرد. اگر اثر بر یک فرد مدنظر است، فرد باید واحد در نظر گرفته شود. اما اگر مدرسه و یا پدیده‌های مرتبط با مناطق جغرافیایی درگیر مسأله هستند، و یا اگر مداخله موردنظر باید الزاماً روی گروهی از افراد انجام شود، آنگاه الزاماً واحد تصادفی‌سازی نباید در سطوح پایین در نظر گرفته شود[۴۴۴]. بنابراین اگرموضوع موردنظر محقق این است که آیا گشت‌زنی متعدد پلیس از جرم و جنایت در یک منطقه جغرافیایی ممانعت می‌کند؟، دفعات مختلف گشت‌زنی باید به مناطق جغرافیایی و نه بلوکهای درون هر منطقه تخصیص داده شود.

در واحدهای مجتمع، ممکن است واحدهای فردی درون مجتمعها مستقل از یکدیگر نباشند، زیرا آنها در معرض تأثیرات عوامل عمومی متعددی غیر از مداخله هستند. به عنوان نمونه، دانش‌آموزان درون یک کلاس درس با یکدیگر صحبت می‌کنند، معلم واحدی دارند، و ممکن است همگی مداخله را در یک زمان واحد از روز دریافت کنند. این وابستگی‌ها منجر به بروز آنچه که سابقاً از آن به نام مسأله واحد تحلیل یاد می‌شد، اما اخیراً بیشتر در قالب مدلهای چندسطحی، و یا مدلهای خطی سلسله‌مراتبی یاد می‌شود، می‌گردد. از آنجا که این کتاب بیشتر بر طرحها تمرکز دارد تا تحلیل، وارد جزئیات مباحث مربوط به آنالیز نخواهیم شد (Feldman, McKinlay, &

[۴۴۴] لانه کردن شـرکت کنندگان در واحدهای سـطح بالاتر می تواند همچنان مشـکلاتی را ایجاد نماید، حتی زمانی که افراد به مداخله ها تخصـیص داده می شوند. برای مثال، اگر افراد مبتلا به سرطان که هر کدام دارای تومورهای متعددی هستند به طور تصادفی به مداخله تخصیص داده شوند اما مداخله به طور مجزا بر روی هر تومور اعمال می شود و عکس العمل تومور به طور مجزا مورد مشاهده قرار می گیرد، این عکس العملها مستقل نیستند.

(Niknian, 1996; Gail et al., 1996; Green et al., 1995; Murray, 1998; Murray et al., 1994; 1996). اما از منظر طراحی آزمایش، استفاده از واحدهای سطح بالاتر مسائل متعددی به همراه دارد.

مطالعاتی که مکرراً از واحدهای سطح بالاتر استفاده می‌کنند، تعداد کمتری واحد برای انجام تصادفی‌سازی در اختیار دارند. موردی را در نظر بگیرید که در آن به دانش آموزان در یک کلاس درس مداخله داده می‌شود، و به دیگران حاضر در کلاس درس دوم که نقش کنترل را دارند، داده نمی شود. شرایط مداخله کاملاً بواسطه کلاسها مخدوش شده است؛ و این باعث می‌شود تا گفتن اینکه آیا تفاوتهای عملکرد در پس‌آزمون در اثر مداخله بوده است، یا مشخصات کلاس درس (مانند کاریزمای معلم، ترکیب دانش آموزان، و یا شرایط فیزیکی کلاس)، عملاً غیرممکن باشد. زمانی که بیش از یک واحد سطح بالا (اما همچنان تعدادی محدود) به شرایط آزمایش تخصیص داده می‌شوند، تصادفی‌سازی می‌تواند منتج به میانگینها، واریانسها، و اندازه‌های نمونه بسیار متفاوت در شرایط آزمایش مختلف شود. با وجود اینکه این موارد مبتلا به مشکلات جدی روایی درونی، و نتایج آماری هستند، در ادبیات موجود به وفور یافت می‌شوند (Simpson, Klar, & Donner, 1995). اینگونه مشکلات اغلب در مورد مطالعات روی مدارس و جوامع رخ می‌دهند؛ زیرا اضافه کردن واحدهای بیشتر بسیار هزینه‌بر است. تخصیص تصادفی واحدهای سطح بالاتر از درون بلوکها و یا طبقه‌ها[۴۴۵] می‌تواند این قبیل مشکلات را کاهش دهد. برای مثال، مک‌کی و همکارانش (McKay et al., 1987) اثرات پنج سطح از برنامه تغذیه، خدمات درمانی، و آموزش را بر توانایی شناختی کودکان دچار سوءتغذیه مزمن در کالی و کلمبیا مورد بررسی قرار دادند. آنها کالی را به ۲۰ منطقه تقریباً همگن تقسیم کردند. سپس آنها مناطق را بر اساس ترتیب استانداردی از نمرات بدست آمده در پایش پیش‌آزمونی رتبه‌بندی کرده، و آنها را به طور تصادفی به یکی از پنج شرایط از بلوکهای پنج‌تایی تخصیص دادند. آزمایش گشت‌زنی پیشگیرانه شهر کانزاس رویه مشابهی را دنبال کرد (Kelling, Pate, Dieckman, & Brown, 1976). محققین ۱۵ منطقه گشت‌زنی را در سه بلوک، که از نظر مشخصه‌های جمعیت‌شناختی مشابه بودند تعبیه کردند؛ آنها سپس مناطق این بلوکها را به طور تصادفی به سه شرایط آزمایشی تخصیص دادند.

برنامه‌ریزی برای اندازه نمونه مناسب و تحلیل طرحهای با واحدهای سطح بالاتر، پیچیده‌تر از معمول است؛ زیرا واحدهای فردی در درون واحدهای مجتمع مستقل نیستند (Bock, 1989; Bryk & Raudenbush, 1992; Bryk, Raudenbush, & Congdon, 1996; H. Goldstein, 1987; Snijders & Bosker, 1999). با فرض برابر بودن تعداد واحدها، توان در طرحهای با واحد سطح بالا همواره پایین‌تر از طرحهای با واحدهای فردی است؛ در نتیجه، تحلیلهای توان ویژه‌ای باید مورد استفاده قرار گیرد[۴۴۶]. بعلاوه، توان بیشتر از طریق افزایش تعداد واحدهای مجتمع ارتقاء پیدا می‌کند (افزایش تعداد کلاسها)، تا با افزایش تعداد افراد داخل واحدها (مثلاً افزودن تعداد

[445] Strata

بیشتر دانش‌آموزان درون کلاسها). بدون تردید از نقطه‌ای به بعد، افزایش بیشتر تعداد دانش‌آموزان اتلاف منابع محسوب می‌شود، بدون آنکه بتواند توان آزمون را ارتقاء دهد. این بستگی به اندازه وابستگی‌های درون‌طبقه‌ای دارد، که از طریق همبستگی‌های دورن‌طبقه ای محاسبه می‌شود.

محدودیت منابع اغلب باعث می‌شود تا محققین از داشتن تعداد واحد سطح بالای لازم که بر اساس تحلیل‌های توان برای انجام تحلیل حساسیت آماری ضروری هستند، بازبمانند. در چنین مواقعی بهتر است مطالعه را به گونه‌ای در نظر بگیریم که انگار شبه‌آزمایش است، و مؤلفه هایی مانند تکرارهای جابجاشونده، و یا پیش‌آزمونهای دوگانه را برای تسهیل استنباط علّی به طرح اضافه کنیم. شدیش و همکارانش (۱۹۸۶) این استراتژی را مورد بحث قرار داده و نمودارهایی را ارائه می‌نمایند. مثلاً در مطالعه کالی، مک‌کی و همکارانش مداخله را به طوری نامتوازن[۴۴۷] میان پنج گروه مداخله اجرا کردند، به گونه‌ای که برخی مداخله را برای کل مدت زمان مطالعه دریافت کردند اما دیگران مداخله را بعدتر دریافت کردند. تمامی واحدها در یک زمان پس‌آزمون را انجام دادند. گزارش محققین که نشان می‌داد که بروز اثرات همزمان با اجرای مداخله در هر یک از گروهها، کمک می‌کند تا تفسیرپذیری نتایج آزمایش افزایش پیدا کند، به رغم آنکه مطالعه مذکور تنها چهار واحد سطح بالا به ازای هر شرایط آزمایش داشت. در نهایت، اندازه‌گیری مشخصات واحدهای سطح بالا کمک می‌کند تا دریابیم به چه میزان این مشخصه‌ها مخدوش‌کننده اثر مداخله بوده‌اند.

محققین بعضی اوقات به طور غیرضروری خود مشکلات واحد تحلیل ایجاد می‌کنند. این حالت زمانی رخ می‌دهد که برای صرفه‌جویی در هزینه‌ها، و کاستن از پیچیدگی‌های لجیستیکی سروکله‌زدن با شرکت‌کنندگان منفرد، مداخلههایی را که قابل اعمال بر افراد بوده را بر گروهها اعمال می‌کنند. با انجام اینکار، محقق وابستگی‌هایی را میان شرکت‌کنندگان درون هر گروه تولید می‌کند. مثلاً یک درمان برای بی‌خوابی را در نظر بگیرید که روی ۵۰ شرکت‌کننده در ۱۰ گروه پنج نفره اجرا می‌شود؛ و علاوه بر این، فرض کنید که درمان می‌توانست به طور انفرادی اجرا شود، زیرا این درمان به لحاظ نظری دارای عناصر بینافردی مانند حمایتهای متقابل بینافردی نیست. تأثیرات گروهی می‌تواند از گروهی به گروه دیگر متفاوت باشد، در نتیجه خروجی نهایی به گونه متفاوتی تحت تأثیر قرار گیرد. بنابراین، در صورتی که سؤال تحقیق امکان اجرای درمان روی افراد را میسر می‌سازد، محققین باید درمان را روی افراد اجرا کنند. در غیر این صورت، عضویتهای گروهی باید در جریان تحلیلها در نظر گرفته شود.

دسترسی محدود تخصیص تصادفی

[447] Staggered

با وجود اینکه تخصیص تصادفی معمولاً بهتر از دیگر عناصر طراحی برای استنباط اینکه تفاوتهای مشاهده‌شده میان گروه درمان و کنترل به دلیل بعضی علّتهاست کارایی دارد، امکان کاربرد آن غالباً محدود است. تخصیص تصادفی تنها زمانی مفید است که محقق به این باور رسیده باشد که استنباط علّی ملکولی ارجحیت دارد. اینگونه استنباطها هدفی معمول در تحقیقات علوم اجتماعی هستند، اما تنها هدف نیستند. در نتیجه تخصیص تصادفی به لحاظ نظری با دیگر اهداف پژوهشی بی‌ارتباط است. مضاف بر این، تخصیص تصادفی تنها بخشی از طرح آزمایشی است، و طرح آزمایشی خود تنها بخشی از یک طرح تحقیق کلی است. طرح آزمایشی برنامه‌ریزی زمانبندی برای مشاهدات، انتخاب درمان و کنترلها، انتخاب مقیاسها، تعیین مشخصات پاسخ‌دهندگان، و شیوه تخصیص واحدها به مداخلهها را در برمی‌گیرد. تخصیص تصادفی تنها با آخرین مورد سروکار دارد؛ بنابراین تخصیص دادن به شیوه تصادفی تضمین‌کننده یک طرح پژوهشی یا آزمایشی نیست.

بنابراین، اگر یک تخصیص تصادفی روی واحدهایی که در تناسب با حوزه نظری و کاربردی پژوهش نیستند انجام شود، سودمندی تحقیق زیر سؤال است، حتی اگر کیفیت استنباطهای علّی بالا باشد. روسی و لیال (Rossi & Lyall, 1976) آزمایش مالیات درآمد منفی نیوجرسی را مورد نقد قرار می‌دهند، زیرا پاسخ‌دهندگان در زمره افراد شاغل فقیر بودند، در حالی که بیشتر حقوقهای کمکی پرداخت‌شده در برنامه ملّی موردنظر، به افراد بیکار فقیر داده می‌شد. در نقدی مشابه، کوک و همکارانش (Cook et al., 1975)، بال و بوگات (ball & Bogatz, 1970) را به خاطر دستکاری سطح پشتگرمی و انگیزش اجتماعی برای (یا در جریان) نشان‌دادن برنامه خیابان کنجد، مورد انتقاد قرار می‌دهند. لارسن (Larson, 1976) آزمایش گشت‌زنی کانزاس را به این دلیل مورد نقد قرار می‌دهد که مقدار گشت‌زنی پلیس حتی در شرایط گشت‌زنی زیاد نیز با میزان متوسط گشت‌زنی در نیویورک برابر نبود، و اینکه تفاوت میان مناطق با گشت‌زنی زیاد و کم در شهر کانزاس به دلیل گذر اتفاقی جوخه ماشینهای پلیس با چراغهای گردان، و آژیر روشن در مناطق تخصیص‌یافته به شرایط کم گشت‌زنی، کاهش می‌یافت. این نقدهای مفید همگی جزئیات آزمایشهای اجتماعی را هدف قرار داده‌اند، و هیچکدام نفس تخصیص تصادفی را به نقد نمی‌کشند. اینگونه نقدها تنها در مواردی که کاربرد تخصیص تصادفی، موجب بروز مشکلاتی شده باشد، برای مطلوبیت تخصیص تصادفی کاربرد دارد. اگرچه این حالت بسیار نادر است.

طرحهایی که با تخصیص تصادفی به کار گرفته می‌شوند

این بخش به مرور طیف وسیعی از طرحهای آزمایشی تصادفی می‌پردازد (جدول ۸.۱ دیگر انواع این طرحها را فهرست کرده است). طرحهایی که ارائه می‌کنیم پرکاربردترین این طرحها در مطالعات میدانی بوده، و بلوکهای اولیه‌ای را تشکیل می‌دهند که با آنها می‌توان طرحهای پیچیده‌تری ساخت (Fleiss, 1998; Keppel, 1991; Kirk, 1982; Winer, Brown, & Michels, 1991). در این قسمت از همان علائمی برای نشان دادن طرحها استفاده می‌کنیم که در فصلهای گذشته بکار گرفته شد. البته حرف R در ابتدای هر طرح نشان‌دهنده آن است که گروه

نشان داده‌شده در آن خط، از طریق تخصیص تصادفی شکل گرفته است. R را در ابتدای هر خط قرار می‌دهیم؛ اگرچه تخصیص تصادفی می‌تواند قبل یا بعد از پیش‌آزمون انجام شود، بنابراین مکان R بسته به زمان انجام تخصیص تصادفی متغیر خواهد بود.

طرح پایه

آزمایش تصادفی پایه حداقل به دو شرایط، تخصیص تصادفی واحدها به شرایط، و پس‌آزمون اندازه‌گیری واحدها نیاز دارد. ساختار این طرح را می‌توان به صورت زیر نشان داد:

R X O

R O

نمونه‌ای خوب از کاربرد این طرح با یک مداخله منفرد و یک گروه کنترل، آزمایش سالک پولیو[۴۴۸] در سال۱۹۵۴ است. بیش از چهارصد هزار کودک به طور تصادفی به شرایط دریافت واکسن یا شرایط دریافت دارونما تخصیص داده شدند (Meier, 1972).

جدول ۸.۱: نمودارهای شماتیک از طرحهای تصادفی

[448] Salk Polio

طرح تصادفی پایه که در آن گروه مداخله با کنترل مقایسه می‌شود

R X O

R O

طرح تصادفی پایه که در آن دو مداخله مقایسه می‌شوند

R X_A O

R X_B O

طرح تصادفی پایه که دو مداخله و یک گروه کنترل را با یکدیگر مقایسه می‌کند

R X_A O

R X_B O

R O

طرح گروه کنترل پیش‌آزمون-پس‌آزمون

R O X O

R O O

طرح مداخله جایگزین با پیش‌آزمون

R O X_A O

R O X_B O

مداخلهها و کنترلهای متعدد با پیش‌آزمون

R O X_A O

R O X_B O

R O O

طرح فاکتوریل

R $X_{A_1B_1}$ O

R $X_{A_1B_2}$ O

R $X_{A_2B_1}$ O

R $X_{A_2B_2}$ O

طرح طولی[۴۴۹]

R O . . . O X O O . . . O

R	O . . . O	X	O	O . . . O	

طرح متقاطع[۴۵۰]

R	O	x_A	O	x_B	O
R	O	x_B	O	x_A	O

مسأله کلیدی ماهیت شرایط کنترل است. انتخاب یک نوع خاص از گروه کنترل به این بستگی دارد که محقق مایل به کنترل چه چیزی است. برای مثال در یک شرایط بدون مداخله، اثرات یک بسته مداخله ملکولی (مشتمل بر تمامی عناصر فعال، غیرفعال، مهم و بی‌اهمیت آن) آزمون می‌شود. اگرچه زمانی که علاقمند به بررسی بخشی از این بسته ملکولی باشیم، کنترل باید دربرگیرنده همه اجزاء بسته، بغیر از بخش موردنظر باشد. برای مثال در مطالعات دارویی، محققین اغلب مایل‌اند اثرات محتویات دارویی فعال در داروها را از اثر مابقی اجزاء بسته (چیزهایی مانند بلعیدن قرص یا داشتن تماس با پرسنل کادر درمان) تفکیک نمایند. یک کنترل دارونما اینکار را برای آنها انجام می‌دهد. در این نوع کنترل پرسنل قرصی خنثی و بی‌اثر را به بیماران می‌دهند اما اینکار را به شیوه‌ای کاملاً شبیه به شرایطی که داروی واقعی به بیماران داده می‌شود، انجام می‌دهند (مابقی عناصر دو شرایط یکسان است) (Beecher, 1955).

گونه‌های مختلفی از گروههای کنترل وجود دارد(Borkovec & Nau, 1972; Garber & Hollon, 1991; International conference on Harminization, 1999; Jacobson & Baucom, 1977; Kazdin & Wicoxon, 1976).

برای مثال، می‌توان به مواردی همچون کنترلهای بدون مداخله، کنترلهای دوز-پاسخ، کنترلهای لیست انتظار، کنترلهای انتظارات[۴۵۱] و یا کنترلهای فقط-توجه[۴۵۲] اشاره داشت. اما در تمامی موارد سوال همواره این است که «چه چیزی را کنترل می‌کنیم؟» برای مثال، رسی و لیال (Rossi & Lyall, 1976, 1978) آزمایش مالیات منفی نیوجرسی را به این دلیل مورد نقد قرار می‌دهند که گروه کنترل نه تنها از نظر دریافت یا عدم دریافت مداخله مربوطه، بلکه از نظر دریافت خدمات اداری بسیار کمتر نیز با گروه مداخله تفاوت دارد.

دو گونه طرح مختلف بر اساس طرح پایه

یک واریته با جایگزین کردن دو مداخله x_B و X_A به جای X و جای خای در نمودار قبلی بدست می آید:

[۴۴۹] Longitudinal
[۴۵۰] Crossover
[۴۵۱] Expectancy
[۴۵۲] Attention-only

R X_A O

R $\mathrm{x_B}$ O

مثلاً اگر x_A یک مداخله جدید و خلاقانه باشد، $\mathrm{x_B}$ اغلب یک مداخله استاندارد با کارایی شناخته‌شده است. در اینجا سوال علّی به این صورت خواهد بود که، «اثر نوآوری، در مقایسه با آنچه در اثر دریافت مداخله استاندارد از سوی واحدها رخ می‌داد، چیست؟» این طرح در صورتی به خوبی عمل می‌کند، که سابقه پر و پیمانی در زمینه مقایسه مداخله استاندارد موردنظر با شرایط کنترل بدون‌مداخله وجود داشته باشد. اما اگر وجود نداشته باشد، و اگر افراد دریافت‌کننده x_A در پس‌آزمون، متفاوت از دریافت‌کنندگان $\mathrm{x_B}$ نباشند، محقق نمی‌تواند بفهمد که آیا هر دو مداخله به یک میزان کارا بوده‌اند یا نه. در این حالت، وجود یک گروه کنترل کمک‌کننده خواهد بود:

R X_A O

R $\mathrm{x_B}$ O

R O

این طرح در بوستون برای مطالعه اثرات یک پروژه مسکن اجتماعی که برای ارتقاء نوع محله‌هایی که خانوارهای فقیر در آنها زندگی می‌کردند، بکار گرفته شد (Katz , Kling, & Liebman, 1997; Orr, 1999). خانواده‌های فقیر در مداخله A کوپنهای مسکن دریافت می‌کردند که تنها برای استفاده در مناطق با سطح پایین فقر مناسب بود؛ به گونه‌ای که در صورت نقل‌مکان، به مناطق بهتری می‌رفتند. خانواده‌های حاضر در شرایط مداخله B، کوپنهایی برای استفاده در همه مناطق را دریافت می‌کردند، به طوریکه می‌توانستند به مناطق با سطح فقر بالاتر و یا پایین‌تر بروند. خانوارهای گروه کنترل هیچ کوپنی دریافت نمی‌کردند.

ریسکهای این طرح بواسطه عدم وجود پیش‌آزمون

در جایی که انجام پیش‌آزمون می‌تواند اثر حساسیت‌زایی ناخواسته داشته باشد، حذف پیش‌آزمون یک انتخاب درست است. همچنین در جایی که نمی‌توان داده‌های پیش‌آزمونی جمع‌آوری کرد (مانند برخی مطالعات در حوزه ارتقاء شناختی نوزادان)، انجام پیش‌آزمون جداً غیرکاربردی است (مانند مصاحبه‌های بسیار گران‌قیمت و زمانبر پزشکان با بیماران)، و یا مقدار پیش‌آزمون ثابت است (مانند مطالعات مرگ و میر، که در آن تمامی بیماران در ابتدای بیماری زنده هستند)، لازم است پیش‌آزمون حذف شود. در غیر این موارد، حذف پیش‌آزمون عموماً خطرناک است، خصوصاً اگر احتمال ریزش در طول مطالعه وجود داشته باشد. در واقع برخی محققین ضرورت وجود پیش‌آزمون را یکی از درسهای اصلی آموخته شده از ۲۰ سال انجام آزمایشهای اجتماعی می‌دانند

(Haveman, 1987). ریزش اغلب در مطالعات میدانی اتفاق می‌افتد، و غالباً محقق باید بررسی کند که (۱) آیا افرادی که مطالعه را ترک کرده‌اند، متفاوت از افرادی بوده‌اند که در مطالعه مانده‌اند، و (۲) آیا افراد خارج‌شده از یک شرایط خارج شده‌اند، و با افرادی که از شرایط دیگر خارج شده‌اند، متفاوت‌اند یا نه. اطلاعات قبل از اعمال مداخله – ترجیحاً روی همان متغیر وابسته‌ای که در پس‌آزمون مورد استفاده قرار گرفته - تا حد زیادی می‌تواند برای پاسخ به این سوالات یاری‌رسان باشد.

به طور قطع، ریزش در مطالعات میدانی اجتناب ناپذیر نیست. در آزمایشهای پزشکی روی رویه‌های جراحی که اثرات بلافاصله دارند، مداخله بسیار سریعتر از آن اتفاق می‌افتد که امکان ریزش بوجود بیاید؛ و مراقبتهای بعد از عمل و مدارک پزشکی مربوط به سوابق به اندازه کافی دقیق و مطلوب هستند، که پس‌آزمونها و مشاهدات پیگیری[۴۵۳] در دسترس است. نمونه‌ای از این مطالعات، پژوهش تیلور و همکارانش (Taylor et al., 1978) بر روی نرخ کوتاه مدت مرگ و میر در میان ۵۰ بیمار دچار سکته قلبی بود، که در جریان احیاء قلبی به طور تصادفی به شرایط دریافت فشار دستی قفسه سینه یا فشار مکانیکی قفسه سینه تخصیص داده شده بودند، می‌پرداخت. مداخله در طول مدت زمان یک ساعت آغاز شده، و خاتمه می‌یافت؛ بیماران دچار سکته قلبی نمی‌توانستند بلند شوند و بیمارستان را ترک کنند، و مقادیر مربوط به متغیر وابسته به سرعت و سهولت اندازه‌گیری و جمع‌آوری می‌شد. دومین موقعیت با حداقل ریسک ریزش، حالتی است که در آن نتایج به موضوعی مرتبط باشد که به طور عمومی اجبار ثبت وجود داشته باشد. برای مثال، در هر دو آزمایش LIFE (بیمه زندگی برای مجرمان آزاد شده) و TARP (کمک سنتی به زندانیان آزاد شده)، متغیر اصلی وابسته، دستگیری بود که اطلاعات مربوط به آن در سوابق عمومی ثبت‌شده برای تمامی زندانیان وجود داشت (Rossi, Berk, & Lenihan, 1980). اگرچه به طور کلی، ریزش از شرایط در غالب آزمایشهای میدانی رخ می‌دهد و پیش‌آزمونها برای روشهایی که برای مقابله با ریزش در فصل ۱۰ مورد بحث قرار خواهند گرفت، حیاتی هستند.

طرح گروه کنترل پیش‌آزمون- پس‌آزمون

اضافه کردن پیش‌آزمونها به طرح تصادفی پایه بسیار توصیه شده است:

R O X O

R O O

یا اگر تخصیص تصادفی بعد از پیش‌آزمون اتفاق بیافتد:

O R X O

[453] Follow up

O R O

این طرح را می‌توان معمول‌ترین طرح آزمایشهای میدانی دانست. مزیت ویژه این طرح در توانایی تقویت‌شده آن برای مقابله با ریزش، به عنوان تهدیدی برای روایی درونی است. دومین مزیت این است که این طرح امکان برخی تحلیلهای آماری خاص که توان آماری را برای رد فرض صفر افزایش می‌دهند را فراهم می‌آورد (Maxwell & Delaney, 1990). مکسول (S. E. Maxwell, 1994) بر این باور است که تخصیص ۷۵٪ از ارزیابی به پس‌آزمون و ۲۵٪ به پیش‌آزمون، اغلب انتخاب خوبی برای حداکثر کردن توان این طرح است. مکسول و همکارانش (Maxwell et al., 1991) مزایا و معایب ANCOVA ای که در آن از پیش‌آزمون به عنوان یک همتغیر[۴۵۴] استفاده می‌شود و ANOVA اندازه‌گیری مکرر را به عنوان روشی برای افزایش توان مورد بحث قرار می‌دهند.

اگرچه تلاش محقق باید بر این باشد که مقیاسهای پیش‌آزمون و پس‌آزمون یکسان باشند، اما این مسأله الزاماً ضروری نیست. برای مثال، در تحقیقی بر روی رشد کودکان، آزمون برای بچه‌های ۸ ساله باید به طور معناداری متفاوت از آزمون سنجش بچه‌های ۳ ساله باشد. اگر پیش‌آزمون و پس‌آزمون سازه تک‌بعدی یکسانی را ارزیابی کنند، نظریه آزمون لجیستسک بعضی اوقات می‌تواند برای یکسان‌سازی آزمونها مورد استفاده قرار گیرد، البته اگر آن آزمونها محتوی مشترکی داشته باشند (شبیه آنچه که مک کی و همکارانش (McKey et al., 1978) در مطالعه کالی در مورد تغییرات در توانایی شناختی در ۳۰۰ کودک ۳۰ تا ۴۸ ماهه انجام دادند).

طرح مداخلههای جایگزین با پیش‌آزمون

اضافه کردن پیش‌آزمون همچنین زمانی که مداخلههای متفاوتی در حال مقایسه هستند توصیه می‌شود:

R O X_A O

R O X_B O

اگر پیش‌آزمون تفاوتی میان گروهها نشان ندهد، محقق می‌تواند نمرات پیش‌آزمون و پس‌آزمون را بررسی کند تا ببیند که آیا هر دو گروه ارتقاء پیدا کرده‌اند، و یا اینکه هیچکدام ارتقاء پیدا نکرده‌اند[۴۵۵]. این طرح آزمایشی خصوصاً در مواقعی که ملاحضات اخلاقی برای مقایسه گروه مداخله و کنترل وجود دارد، مفید است؛ مثلاً در تحقیقات پزشکی که در آنها تمامی بیماران باید مورد درمان قرار گیرند. این طرح همچنین زمانی که بعضی مداخلهها به عنوان درمان استاندارد شناخته می‌شوند، و دیگر درمانها در مقایسه با آنها باید سنجیده شوند، کاربرد دارد. مقایسه‌ها با این درمان استاندارد، کاربردهای ویژه‌ای برای تصمیم‌گیریهای بعدی دارد.

---

[454] Covariate

[۴۵۵] اینجا و جاهای دیگر، منظور این نیست که تغییر نمرات به عنوان مقیاسی برای نشان دادن این ارتقاء مطلوب است. ANCOVA معمولاً بسیار قویتر است و ملاحضات مرتبط با خطی بودن و همگونی رگرسیون کمترین اهمیت را برای نمرات تغییر به شیوه ای نظیر ANCOVAدارند.

مداخلهها و کنترلهای متعدد با پیش‌آزمون

آزمایش تصادفی با پیش‌آزمون می‌تواند شامل یک گروه کنترل و چندین گروه مداخله باشد:

R	O	X_A	O
R	O	X_B	O
R	O		O

مطالعه بلوم (Bloom, 1990) درباره خدمات شغل‌یابی مجدد برای کارکنان تعدیل‌شده از این طرح آزمایشی استفاده می‌کند. بیش از ۲۰۰۰ کارگر بیکارشده به صورت تصادفی به یکی از موقعیتهای کمک برای جستجوی کار، کمک برای جستجوی کار بعلاوه آموزشهای حرفه ای، و یا بدون مداخله تخصیص داده شدند. در نظر داشته باشید که اولین مداخله تنها یک بخش از مداخله موقعیت دوم را شامل می‌شد، این باعث می‌شد تا بتوان فهمید کدام بخش از مداخله بیشترین اثر را داشته است. بعضی اوقات از این طرح با عنوان مطالعات اوراق کردن[۴۵۶] یاد می‌شود. اگرچه مطالعه‌ای که به عنوان نمونه ذکر شد اوراق کامل حساب نمی‌شود، چون شرایط تنها آموزش را نداشت. روشن است که محدودیت منابع و سختی‌های لجیستیکی غالباً محقق را از بررسی موقعیتهای خیلی زیاد منع می‌کند؛ برای هر کدام از شرایط لازم است تا تعداد زیادی شرکت‌کننده داشته باشیم تا بتوانیم به خوبی آزمون کنیم. و اینکه تمامی موقعیتها ارزش بررسی کردن ندارند؛ خصوصاً شرایطی که احتمال بکارگیری در سیاستگذاری را ندارد.

این طرح آزمایشی را می‌توان بسط داده تا بیش از دو مداخله را دربرگرفته، و یا بیش از یک گروه کنترل داشته باشد. برنامه تحقیقاتی موسسه ملی سلامت روان برای درمان مشارکتی افسردگی مثال خوبی از این گونه بسط دادن به حساب می‌آید (NIMH-TDCRP; Elkin, Perloff, Hadley, & Autry, 1985; Elkin et al., 1989; Imber et al., 1990). در این مطالعه ۲۵۰ بیمار افسرده به طور رندم به یکی از شرایط دریافت درمان‌شناختی رفتاری، رواندرمانی بینافردی، و یا دارونما بعلاوه مدیریت کلینیکی تخصیص داده شدند.

این طرح همچنین برای تنوع بخشیدن به متغیرهای مستقل در مجموعه‌ای از سطوح افزایش‌یابنده (برخی اوقات به آنها مطالعات پارامتریک و یا دوز-پاسخ نیز گفته می‌شود) مورد استفاده قرار می‌گیرند. برای مثال، در آزمایش کمک هزینه تقاضای خانه، خانوارها به شرایط دریافت کمک هزینه‌های ۰٪، ۲۰٪، ۳۰٪، ۴۰٪، ۵۰٪ و یا ۶۰ ٪ اجاره تخصیص داده شدند (Friedman & Weinberg, 1983). آزمایش بیمه سلامت به طور تصادفی خانوارها را به یکی از شرایط برنامه‌های بیمه‌ای که نیازمند پرداخت ۰٪، ۲۵٪ ، ۵۰٪ ، و یا ۹۵٪ از ۱۰۰۰ دلار اولیه برای خدمات پوشش بیمه‌ای بود، می‌نمود (Newhouse, 1993). هرچه سطح بالاتری از مداخله اعمال می‌شود، ارزیابی

[456] Dismantling

شکل کارکردی اثر-دوز بطور سطحی‌تری انجام خواهد شد. داشتن طیف وسیعی از سطوح مداخله این امکان را برای مطالعه فراهم می‌آورد که بتوان اثراتی را تشخیص داد، که در صورت وجود تنها دو سطح مداخله (که آنقدر هم قدرتمند نبودند تا اثری متفاوت داشته باشند)، ممکن بود از نظر مخفی بمانند. برای مثال، مطالعه کالی و کلمبیا یک مداخله ترکیبی آموزشی، تغذیه‌ای و درمانی را در چهار سطح افزایش‌یابنده ۹۹۰ ساعت، ۲۰۷۰ ساعت، ۳۱۳۰ ساعت و ۴۱۷۰ ساعت اجرا کرد (McKay et al., 1978). در کمترین دوز - که خود چیزی در حدود یک سال کامل طول کشیده، و از نظر بسیاری از محققین می‌توانست دوز ماکزیمم تلقی شود- اثر تقریباً غیرقابل تشخیص است. اما مک‌کی و همکارانش (Mckey et al., 1978) همچنان اثرات مداخله را مشاهده کردند، چون آنها طیف وسیعی از دوزهای بالاتر را در مطالعه گنجانده بودند.

طرحهای فاکتوریل

این طرحها دو یا بیشتر متغیر مستقل (که به آنها فاکتور گفته می‌شود) را در حداقل دو سطح برای هر یک، مورد استفاده قرار می‌دهند (شکل ۸.۱). برای مثال، فرض کنید بخواهیم اثر یک ساعت راهنمایی تحصیلی در هفته (فاکتورA، سطح۱) را با اثر چهار ساعت راهنمایی (فاکتور A، سطح ۲) با یکدیگر مقایسه کرده، و همچنین راهنمایی انجام شده توسط یک همسال (فاکتور B، سطح ۱) را با راهنمایی انجام‌شده توسط یک فرد بالغ (فاکتورB، سطح ۲) مقایسه نماییم. اگر مداخله در طرحی فاکتوریلی ترکیب شوند، چهار گروه یا سلول شکل می‌گیرد: (۱) یک ساعت راهنمایی تحصیلی از طریق یک همسال (سلول $X_{A_1B_1}$)، (۲) یک ساعت راهنمایی تحصیلی از طریق یک فرد بالغ (سلول $X_{A_1B_2}$)، (۳) چهار ساعت از طرف یک همسال (سلول $X_{A_2B_1}$)، و (۴) چهار ساعت توسط یک فرد بالغ (سلول $X_{A_2B_2}$). این طرح را غالباً طرح فاکتوریل ۲x۲ (دو در دو) نامیده، و در این کتاب آن را به صورت زیر نشان می‌دهیم:

R	$X_{A_1B_1}$	O
R	$X_{A_1B_2}$	O
R	$X_{A_2B_1}$	O
R	$X_{A_2B_2}$	O

این منطق را می‌توان برای طرحهایی با بیش از دو فاکتور نیز بسط داد. اگر یک عامل سوم که در آن راهنما با روشهای اثربخش آموزش دیده یا ندیده باشد، در این حالت ما یک طرح ۲x۲x۲ خواهیم داشت که دارای ۸ سلول خواهد بود. سطوح فاکتورها می‌تواند شامل شرایط کنترل هم باشد، برای مثال، می‌توان یک شرایط بودن

راهنمایی را به فاکتور A اضافه کرد. این کار تعداد سطوح A را افزایش داده به طوریکه یک طرح ۳x۲x۲ با ۱۲ سلول خواهیم داشت. این کار را می‌توان برای فاکتورهای بیشتر، و سطوح بیشتر نیز به همین شیوه انجام داد. طرحهای فاکتوریل دارای سه مزیت بزرگ هستند:

- اغلب تعداد واحدهای (شرکت‌کنندگان) کمتری نیاز دارند؛
- امکان آزمون ترکیبی از مداخلهها را به صورت ساده‌تری فراهم می‌آورند؛
- امکان آزمون برهم‌کنشها[۴۵۷] را فراهم می‌آورند.

اول، این طرحها اغلب امکان داشتن نمونه‌ای کوچکتر از آن‌چیزی که در صورت بکارگیری طرحی غیر از این طرحها مورد نیاز بود، را میسر می‌سازند[۴۵۸]. یک آزمایش برای آزمون تفاوت میان راهنمایان همسال در مقابل بالغ، احتمالاً نیازمند ۵۰ شرکت‌کننده به ازای هر یک از شرایط بود، همینطور برای مقایسه ۱ ساعت و ۴ ساعت نیز، به ۵۰ نفر به ازای هر شرایط نیاز داشتیم؛ این برابر خواهد بود با ۲۰۰ شرکت‌کننده در مجموع. در یک طرح فاکتوریل، کمتر از ۲۰۰ شرکت‌کننده نیاز خواهیم داشت زیرا هر شرکت‌کننده یک وظیفه دو برابری را به عهده می‌گیرد، با توجه به اینکه به طور همزمان در معرض هر دو مداخله قرار دارند.
دوم، طرحهای فاکتوریل به محقق اجازه می‌دهد تا این موضوع را که آیا ترکیبی از مداخلهها، اثربخش‌تر از یک مداخله است را آزمون کنند. فرض کنید که یک محقق دو آزمایش انجام می‌دهد، یکی آزمایشی که در آن برای درمان سردردهای میگرنی به شرکت‌کنندگان آسپرین یا دارونما داده می‌شود، و آزمایش دومی که در آن به همان منظور، اثر بیوفیدبک و دارونما مورد آزمون قرار می‌گیرد. این دو آزمایش هیچ اطلاعاتی در مورد اینکه اگر بیوفیدبک و آسپیرین با هم تجویز شوند، ارائه نمی‌کند. طرح فاکتوریل در مورد اثرات تنها آسپیرین، تنها بیوفیدبک، بیوفیدبک بعلاوه آسپیرین، و یا بدون دارو اطلاعات ارائه می‌نماید.

فاکتور B

		سطح ۱	سطح ۲	
فاکتور A	**سطح ۱**	سلول A_1B_1	سلول A_1B_2	**میانگین ردیف برای A_1**
	سطح ۲	سلول A_2B_1	سلول A_2B_2	**میانگین ردیف برای A_2**
		میانگین ستون برای B_1	**میانگین ستون برای B_2**	

[۴۵۷] Interaction

[۴۵۸]

شکل ۸.۱. اصطلاحات و شیوه نمایش طرح فاکتوریل

سوم، آزمایشهای فاکتوریل برهم‌کنشها میان فاکتورها را آزمون می‌کند (Abelson, 1996; D. Meyer, 1991; Petty, Fabrigar, Wegener, & Priester, 1996; Rosnow & Rosenthal. 1989. 1996). مداخلهها اثرات اصلی ایجاد می‌کنند؛ مثل اثر اصلی آسپیرین در مقایسه با قرص دارونما برای کاهش سردرد. اثرات اصلی اثرهای متوسطی هستند، که اگر برای مثال برخی انواع سردردها به آسپیرین به خوبی جواب می‌دهند، اما دیگر سردردها جواب نمی‌دهند، می‌توانند گمراه‌کننده باشند. برهم‌کنشها زمانی رخ می‌دهند که اثرات مداخله ثابت نیستند، بلکه به ازای سطوح متفاوت دیگر فاکتورها، تغییر می‌کنند؛ برای مثال اینکه آسپیرین سردردهای ناشی از تنش را به میزان زیاد، اما سردردهای میگرنی را به اندازه‌ای اندک کاهش می‌دهد. در اینجا مداخله موردنظر (آسپیرین) با یک متغیر تعدیلگر[۴۵۹] (نوع سردرد) برهم‌کنش دارد. کلمه تعدیلگر به فاکتور دومی اشاره دارد که با یک مداخله برهم‌کنش دارد (اثر آن را تعدیل می‌کند). همین منطق برای طرحهای با سه یا بیشتر فاکتور مصداق دارد. البته تفسیر تعاملات سطح بالاتر دشوار است.

تعاملات را سخت‌تر از اثرات اصلی می‌توان تشخیص داد (Aiken & West, 1991; Chaplin, 1991; 1997; Cronbach & Snow, 1977; Fleiss, 1986). در نتیجه، زمانهایی که بررسی برهم‌کنشها مدنظر است، نمونه‌های بزرگ، با تحلیلهای توان آماری مناسب ضروری است[۴۶۰]. بدون شک، برخی محققین بر این باورند که برهم‌کنشهای پیش‌بینی‌شده آنقدر در خلق نظریه‌های علمی جدید مهم هستند که آزمون آنها با خطای نوعی اولی در سطحی بالاتر از معمول، همچنان توجیه‌پذیر است (Meehl, 1978; Platt, 1964; Smith& Sechrest, 1991; Snow, 1991). اگر آزمون یک برهم‌کنش پیش‌بینی‌شده موردنظر باشد، گرفتن نمونه‌های بیشتر از نیاز، از مشاهداتی که از نظر متغیر تعدیلگر مقادیر خیلی بالا یا خیلی پایین دارند، می‌تواند آزمون قویتری (و همچنان بدون سوگیری) از برهم‌کنشها فراهم نماید. اگرچه، تخمین ضعیفتری از واریانس کل احصاءشده توسط متغیرهای موردنظر ارائه خواهد داد. آزمون بررسی برهم‌کنش را می‌توان با نمونه غیروزنی[۴۶۱] انجام داد، و آزمون واریانس کل را می‌توان با نمونه‌ای که به منظور انعکاس ترکیب جمعیت موردنظر وزنی شده است، انجام داد (McClelland, 1985). این موردی ویژه از نظریه طرح بهینه (Atkinson, 1985) است، که می‌تواند به انتخاب

---

459 Moderator

۴۶۰ تعاملات می توانند ترتیبی (ordinal) (زمانی که میانگین سلولها را روی نمودار می برید خطهای بدست آمده یکدیگر را قطع نمی‌کنند) یا غیرترتیبی (خطهای بدست آمده یکدیگر را قطع می کنند) باشند. آزمون اثرات تعاملات ترتیبی مکرراً توان کمتری نسبت به اثرات اصلی دارند، اما آزمون های تعاملات غیر ترتیبی اغلب قویتر از آزمون هر کدام از اثرات اصلی هستند. روزنو و روزنتال (۱۹۸۹) این مساله را که در هنگام وجود تعاملات، کدامیک از خط ها باید و یا نباید قطع شوند را توضیح می دهند.

461 Unweighted

سطوح مداخله، و ترکیبهایی که قدرت طرح را برای تشخیص پارامترهای موردتوجه نظریات یا سیاستها حداکثر می‌کنند، کمک نماید.

در هنگام استفاده از طرح فاکتوریل، لازم نیست محقق به طور واقعی واحدها را به تمامی ترکیبهای ممکن از فاکتورها تخصیص دهد؛ اگرچه، وجود خانه‌های خالی می‌تواند توان آزمون را کاهش دهد. این کار می‌تواند منابع محقق را برای آزمون آن دسته از ترکیبهای درمانی که از نظر تئوریک و یا کاربردی جذابیتی ندارند، هدر دهد. برای مثال، آزمایش مالیات منفی نیوجرسی پیشنهاداتی را برای مقابله با فقر و رفرم رفاه مورد بررسی قرار می‌داد (علی‌الخصوص اثر ترکیبی دو متغییر مستقل سطح گارانتی و نرخ مالیات را) (Kershaw & Fair, 1976). سطح گارانتی پولی است که به خانوارهای فقیر پرداخت می‌شود، اگر هیچ درآمد دیگری نداشته باشند. این متغیر به صورت ۵۰٪، ۷۵٪، ۱۰۰٪، و یا ۱۵٪ درصد خط فقر تعریف می‌شود. نرخ مالیات نرخی است که بر اساس آن درآمد گارنتی‌شده بر مبنای افزایش درآمد خانوار، به میزان ۳۰٪، ۵۰٪، ۷۰٪ کاهش داده می‌شود. بنابراین طرح آزمایش یک طرح آزمایشی فاکتوریل ۳x۴ است که می‌توانست شرکت‌کنندگان را به ۱۲ سلول متفاوت تخصیص دهد. اگرچه، محققین ۷۲۵ شرکت‌کننده را به ۸ سلولی تخصیص دادند که بیش از حد هزینه‌بر نبوده، و از نظر سیاسی و سیاستگذاری انجام‌شدنی و توجیه‌پذیر بودند. سلولهای خالی می‌توانند تحلیل داده‌ها را پیچیده کنند، اما انعطاف‌پذیری این گزینه اغلب بیشتر از پیچیدگیهای آن است. این طرح مثالی است از یک طرح فاکتوریل بخشی، که امکان تخمین برخی برهم‌کنشهای سطح بالاتر را فراهم می‌آورد، حتی زمانی که طرح فاکتوریل کامل به کار گرفته نمی‌شود (Anderson & Mc Lean, 1984; Box, Hunter, & Hunter, 1978; West, Aiken, & Todd, 1993).

طرح های لانه‌گزیده و متقاطع

در طرح متقاطع، هر سطح از هر فاکتور در معرض (در تقاطع با) تمامی سطوح دیگر فاکتورها قرار می‌گیرد. برای مثال، در یک آزمایش آموزشی، اگر برخی دانش‌آموزان در هر کلاس درس در معرض مداخله، و برخی دیگر در گروه کنترل قرار بگیرند، می‌توان گفت که فاکتور مداخله در تقاطع با کلاس درس است. در یک طرح لانه‌گزیده، برخی سطوح یک فاکتور در معرض تمامی سطوح دیگر فاکتورها قرار نمی‌گیرند. برای مثال، زمانی که برخی کلاسهای درس در گروه مداخله قرار می‌گیرند، و نه در گروه کنترل، کلاسها در درون شرایط مداخله لانه گزیده‌اند. طرحهای متقاطع آزمونهای آماری مخدوش‌نشده برای اثرات اصلی و اثرات برهم‌کنشی بدست می‌دهند، اما طرحهای لانه‌گزینه ممکن است اینطور نباشند. این تفاوت میان طرحهای لانه‌گزیده و متقاطع به ویژه در هنگام وجود واحدهای سطح بالاتر (مانند مدارس، بیمارستانها، محلهای کار) بروز می‌نماید. مسأله این است که طرحهای متقاطع تخمینهای مجزای آماری از اثرات واحدهای سطح بالاتر، شرایط مداخله، و برهم‌کنشهای میان آنها بدست می‌دهد، اما احتمال مشکلاتی مانند انتشار مداخله را افزایش می‌دهد. هر محقق

باید پیش از تصمیم برای انتخاب طرح لانه‌گزیده یا متقاطع، ویژگیهای این دو طرح را در زمینه کارایی خاص آنها برای آزمایش موردنظرش بررسی کند.

یکی از معایب طرحهای فاکتوریل

طرحهای فاکتوریل در تحقیقات آزمایشگاهی و در شرایط بسیار کنترل‌شده، مانند شرایط برخی تحقیقات پزشکی معمول هستند. اجرای این طرحها در بسیاری از شرایط آزمایشهای میدانی دشوارتر است. آنها نیازمند کنترلهای دقیقتر در طول ترکیب مداخله ارایه شده به هر یک شرکت‌کننده هستند. اما چنین کنترلی با افزایش فاکتورها و یا سطوح فاکتورها دشوار می‌شود؛ به ویژه اگر هر یک از سلولها معیار واجد شرایط متفاوتی داشته باشد (مانند آنچه در مطالعات دارویی اتفاق می‌افتد، که قوانین درباره اینکه چه کسی چه ترکیب دارویی را می‌تواند دریافت کند، پیچیده است). بعلاوه، بسیاری از تحقیقات میدانی برای ارزیابی کاربرد نوآوریهای جدید در سیاستها انجام می‌شوند. اما توانایی سیاستگذاران برای قانون‌گذاری یا اداره‌کردن برهم‌کنشها، با در نظر گرفتن سنتهای موجود در زمینه کنترلهای محلی و احتیاطهای حرفه‌ای در عرضه خدمات، و با توجه به سختیهای تضمین اینکه مداخلات اجتماعی همانطور که مدنظر بوده اجرا شوند، بسیار محدود است (Pressman & Wildavsky, 1984; Rossi & Wright, 1984). سیاستگذاران غالباً بیشتر علاقمند به استنباطهای کلی در مورد اینکه کدام مداخله کارآمد خواهد بود هستند، و نه استنباهای بسیار جزئی و محلی درباره اثرات ترکیبی خاص از سطوحی خاص از فاکتوری خاص در شرایطی خاص که از طرح‌های فاکتوریل به دست می‌آید.

طرحهای طولی[۴۶۲]

طرحهای طولی مشاهدات متعددی که قبل، بعد و در جریان اعمال مداخله جمع‌آوری شده‌اند، را به مطالعه اضافه می‌کنند، تعداد و زمانبندی هرکدام از این مشاهدات توسط فرضیات مطالعه تعیین می‌شود:

R O . . . O X O O . . . O

R O . . . O X O O . . . O

این طرحها بسیار شبیه مطالعات سری‌زمانی هستند (فصل ۶)، اما این مطالعات مشاهدات پیش‌آزمون و پس‌آزمون بسیار کمتری دارند. طرحهای طولی امکان بررسی و اندازه‌گیری اینکه اثرات چطور در طول زمان تغییر می‌کنند را فراهم می‌آورد. همچنین امکان استفاده از مدلهای منحنی رشد تفاوتهای فردی در پاسخ به مداخله را فراهم آورده، و غالباً قویتر از طرحهایی با تعداد کمتر مشاهدات در طول زمان هستند؛ خصوصاً اگر پنج یا بیشتر موج اندازه‌گیری انجام شده باشد (Maxwell, 1998)، یا زمانی که اندازه نمونه کوچک بوده، و اضافه‌کردن پیش‌آزمون و پس‌آزمون می‌تواند توان آماری را افزایش دهد.

[462] Longitudinal

آزمایشات تصادفی طولی با پیش‌آزمونهای متعدد کمیاب‌اند. برای مثال، بلوم (Bloom, 1990) به طور تصادفی کارکنان تعدیل‌شده را به شرایط سه مداخله و گروه کنترل تخصیص داد؛ مداخلهها برای کمک به این افراد برای پیدا کردن کار طراحی شده بود. بلوم درآمدهای فصلی را در جریان چهار پیش‌آزمون و چهار پس‌آزمون گزارش کرد. پیش‌آزمونها نشان داد که شرکت‌کنندگان در طول یک یا دو فصل قبل از تخصیص به شرایط، کاهش شدیدی در درآمد داشته‌اند. این شاید نشان‌دهنده یک بیکاری کوتاه‌مدت کارگرانی که به سرعت به بازار کار وارد و خارج می‌شوند. بنابراین، اثرات رگرسیونی ممکن است باعث افزایشهایی در تمام گروهها می‌شدند، حتی اگر مداخله‌ها نیز عمل نمی‌کردند. البته شرکت‌کنندگان گروه کنترل نیز پیشرفت کرده بودند، اما نه به اندازه شرکت‌کنندگان گروه مداخله.

استفاده از پس‌آزمونهای متعدد معمول‌تر است. به عنوان نمونه، مطالعه جوانان کمبریج سامرویل در سال ۱۹۳۹ آغاز شد، زمانی که ۶۵۰ نوجوان پسر به صورت تصادفی از جفتهای بلوک‌بندی‌شده به یکی از شرایط دریافت برنامه مشاوره (مداخله)، و یا بدون مداخله تخصیص داده شدند (W. McCord & McCord, 1959; Powers & Witmer, 1951). مطالعه پیگیری ۳۷ سال بعد در سال ۱۹۷۶ انجام شد (J. McCord, 1978). مطالعه‌ای در حال انجام، در یکی از سازمانهای سلامت درصدد آن است که بیماران را تا انتهای عمرشان دنبال نماید. در اینجا پس‌آزمونهای متعدد نشان خواهند داد که آیا اثرات حاصل‌شده از مداخله باقی مانده‌اند، یا اینکه در طول زمان تغییر کرده‌اند. این موضوع بویژه زمانی اهمیت پیدا می‌کند، که نتایج اولیه را تنها چند سال بعد بتوان اندازه‌گیری کرد. مثلاً موفقیت تحصیلی و شغلی کودکان شرکت‌کننده در برنامه آموزش پیش‌دبستانی، و یا درآمد کسب شده در طول عمر کاری در میان افرادی که آموزش ضمن‌خدمت دریافت کرده اند. برخی مواقع مطالعات طولی به طور همزمان نتایج متفاوتی را در طول زمان دنبال می‌کنند، تا اعتبار و روایی زنجیره‌ای مفروض از اثرات علّی را بررسی نمایند. برای مثال، اینکه یک مداخله به کودکان طبقه‌های پایین اقتصادی-اجتماعی کمک کند تا از فقر بیرون بیایند، ابتدا امیدواری و باور آنها نسبت به موفقیت را بهبود خواهد بخشید، که این بر انتظارات آنها اثر خواهد داشت، و متعاقباً موفقیت آنها را در درس دستور زبان تحت تأثیر قرار می‌دهد، و این به آنها کمک می‌کند تا دبیرستان را با موفقیت به پایان برسانند، که این در نهایت منجر به داشتن کاری با درآمد بالاتر در بزرگسالی خواهد بود. در اینجا زمانبندی مشاهدات پیگیری از برنامه زمانبندی رویدادهایی که باید موردمشاهده قرار بگیرند تا بتوان دریافت آیا این زنجیره در جایی دچار انقطاع می‌شود یا نه، پیروی می‌کند. مشکلات عملی آفت مطالعات طولی هستند. اول، با طولانی‌ترشدن دوره‌های پیگیری، ریزش افزایش می‌یابد. برای مثال، شرکت‌کنندگان به مکان دیگری نقل مکان می‌کنند، و یا اینکه از ادامه مطالعه خسته می‌شوند. با این وجود، آنچه که با استفاده از رویه‌های پیگیری از بسیاری از جمعیتها می‌توان بدست آورد، بسیار قابل‌توجه است. دوم، برخی نتایج بدست‌آمده در طولانی مدت (مانند درآمد بدست‌آمده در طول عمر کاری فرد)، با در نظر

گرفتن تکنولوژیهای کنونی، و دسترسی محدود به داده‌های مرتبط غیرممکن به نظر می‌رسد (Boruch & Cecil, 1979). سوم، برخی مواقع محروم کردن گروه کنترل از دسترسی به نوعی از مداخله غیراخلاقی محسوب می‌شود. ضمن اینکه در مطالعات طولی اینکار چندان انجام‌شدنی نیست، زیرا این افراد به نحوی و از طریق دیگر منابع، به درمان موردنظر دست خواهند یافت. نمونه‌ای از تمامی این مشکلات در مطالعه اشنایدر و ویلز (۱۹۸۹) قابل مشاهده است (Snyder & Wills, 1989; Snyder, Wills, & Grady-Fletcher, 1991). این دو محقق به طور تصادفی ۷۹ زوج دچار مشکلات خانوادگی را به یکی از شرایط دریافت خانواده‌درمانی رفتاری (۲۹ نفر)، خانواده‌درمانی بینش-محور (۳۰ نفر)، و یا یک لیست انتظار (گروه کنترل) (۲۰ نفر) تخصیص دادند. در اندازه‌گیریهای پیگیری که یکبار بعد از ۴ ماه، و بار دیگر بعد از ۶ ماه صورت گرفت، محققین تنها مقادیر مربوط به دو گروه را اندازه‌گیری کردند، زیرا گروه کنترل علیرغم آنکه به آنها گفته شده بود که دوره انتظار ممکن است چیزی حدود سه ماه طول بکشد، خسته شده و مطالعه را ترک کرده بودند. علیرغم فوت شرکت‌کنندگان، مشکلات سلامتی و نقل و انتقال محل زندگی، محققین توانستند برای چهار سال از ۵۵ خانوار از ۵۹ خانوار داده پیگیری جمع‌آوری کنند (که نرخ قابل‌توجهی از حفظ شرکت‌کنندگان به حساب می‌آید. اگرچه از دست دادن شرکت‌کنندگان به میزان یک واحد، تقریباً همواره در مطالعات طولی وجود دارد). در نهایت، یک مطالعه پیگیری ۴ ساله بسیار طولانی‌تر از آن چیزی است که بسیاری از مطالعات رواندرمانی مورد استفاده قرار می‌دهند، اما در عین حال، بسیار کوتاه‌تر از آن چیزی است که در مطالعات مربوط به سطح استرس در زندگی زناشویی، یا نرخ طلاق در طول زندگی معمول است. اما حتی در این نمونه بسیار مطلوب از مطالعات طولی نیز اینگونه مشکلات بروز می‌کند.

طرحهای متقاطع[۴۶۳]

آزمایشی را در نظر بگیرید که در آن برخی شرکت‌کنندگان به طور تصادفی به یکی از شرایط دریافت مداخله A یا B تخصیص داده می‌شوند؛ و پس از آن پس‌آزمون اجرا می‌شود. در یک طرح متقاطع، بعد از پس‌آزمون، شرکت‌کنندگان به صورت متقاطع مداخله‌ای را دریافت می‌کنند که در مرتبه قبل دریافت نکرده بودند، و پس از دریافت مداخله دوم، دوباره یک پس‌آزمون انجام می‌شود. در این کتاب، طرح مذکور را به صورت زیر نمایش می‌دهیم:

$$R \quad O \quad x_A \quad O \quad x_B \quad O$$

[463] Crossover designs

$$R \quad O \quad X_B \quad O \quad X_A \quad O$$

برخی اوقات فاصله زمانی میان مداخلهها به درازا می‌کشد، به طوریکه اثرات مداخله اول، قبل از شروع مداخله دوم می‌تواند از بین برود.

این طرح اغلب در تحقیقات پزشکی مورد استفاده قرار می‌گیرد، مانند مواردی که در آن چندین دارو در قالب یک طرح درون‌گروهی به شرکت‌کنندگان داده می‌شود، و متقاطع‌سازی برای موازنه معکوس[۴۶۴] و ارزیابی اثرات ترتیب [مداخلهها] بکار گرفته می‌شود. این طرحها همچنین برای جمع‌آوری اطلاعات علّی بیشتر از مطالعه‌ای که به طور معمول بعد از اجرای پس‌آزمون اول متوقف می‌شد، بکار گرفته می‌شوند. کاربرد این طرحها به هر دلیلی که باشد، طرحهای متقاطع زمانی بیشترین کارایی دارند که مداخلهها بعد از مدت کوتاهی اثر خود را از دست می‌دهند (در غیر این صورت اثرات متعاقب[۴۶۵] رخ خواهد داد)، مداخله به سرعت عمل می‌کند (در غیر این صورت، انجام آزمایش زمان طولانی به طول خواهد انجامید)، و زمانی که شرکت‌کنندگان قادرند و مایلند که هر دو مداخله را انجام دهند، حتی اگر مداخله اول مشکل را برطرف کرده باشد. اگر تحلیلها تعاملی را میان مداخله‌ها و ترتیب اعمال آنها نشان دهد، آن زمان اثر دور دوم مداخله را بدون در نظر گرفتن اثرات ترتیب مداخله‌ها نمی‌توان تفسیر کرد؛ اگرچه دور اول مداخله همچنان قابل‌تفسیر است (دقیقاً مانند زمانی که طرح متقاطع انجام نشده بود).

مناسب‌ترین شرایط برای انجام تخصیص تصادفی

این بخش (و جدول ۸.۲) شرایطی را به تصویر می‌کشند که احتمال انجام موفقیت‌آمیز یک آزمایش تصادفی را افزایش می‌دهد.

زمانی که تقاضا بر عرضه پیشی می‌گیرد

زمانی که تقاضا برای نوعی از خدمات، از عرضه بیشتر می‌شود، تصادفی‌سازی می‌تواند منطقی معتبر برای توزیع منصفانه خدمات باشد. برای مثال، دانفورد (Dunford, 1990) آزمایشی را در مورد اثرات یک برنامه اشتغال تابستانه جوانان تشریح می‌کند. در ابتدا، پرسنل برنامه دانش‌آموزان را به طور تصادفی به شرایط اشتغال و غیراشتغال تخصیص دادند. اگرچه، آنها متوجه شده بودند که تعداد شغلهای موجود بسیار کمتر از تعداد متقاضیان است، و به این تصمیم رسیده بودند که تخصیص تصادفی آن شغلها منصفانه است. آنها بعدها گزارش کردند که ماهیت بدون‌سوگیری تصادفی‌سازی به آنها کمک کرده تا به یک گروه از منتقدین نشان دهند که ورود به برنامه اشتغال تبعیضی بر علیه یا به نفع جوانان متعلق به گروههای اقلیت دربرنداشته است. به همین

[464] Counterbalance
[465] Carryover effect

طریق، قانون تلفیق کلیات بودجه به ایالات اجازه می‌داد تا برای رفرمهای حوزه رفاه آزمایشهایی با رویکردهای نو و بدیع بکار بگیرند. بسیاری از ایالات مایل بودند این کار را انجام دهند، اما ایالات معدودی توانستند هزینه اجرای برنامه هایی را که می‌توانست برای عموم مردم دریافتهای نقدی اجرا کند، را تحمل کنند. در اینجا نیز تخصیص تصادفی به عنوان مکانیسمی منصفانه برای توزیع خدمات یک آزمایش پذیرفته شد. در نهایت، در برنامه انتخاب والدین میلواکی، هنگامی که تعداد متقاضیان برای یک مدرسه و یا یک سال تحصیلی بیشتر از ظرفیت موجود بود، کاربرد کوپن مدارس از طریق انتخاب تصادفی شرکت‌کنندگان مورد آزمون قرار گرفت (Rouse, 1998).

زمانی که تقاضا از عرضه بیشتر است، متقاضیانی که به شرایط کنترل تخصیص داده شده‌اند دوباره متقاضی دریافت مداخله می‌شوند. آزمایشگران باید درباره اینکه آیا این افراد چنین حقی دارند یا نه، به روشنی تصمیم‌گیری کنند؛ و اگر دارند، آیا متقاضیانی که مجدداً مراجعه کرده‌اند، اولویتی نسبت به متقاضیان جدید دارند یا نه. بعضی اوقات به دلیل مسائل اخلاقی نمی‌توان حق دوباره تقاضا دادن را از افراد سلب نمود؛ مانند مورد یک بیمار رواندرمانی که به لیست انتظار تخصیص داده شده، اما به کرّات دچار علایم شدید می‌شود و یا یک دریافت‌کننده کمک‌هزینه‌های دولتی که به طور قانونی حق تقاضای مجدد برای برنامه آموزش ضمن‌خدمت را دارد. این نکته بسیار حائذ اهمیت است که برای بدست‌آوردن پشتیبانی همه افراد دخیل در آزمایش در مورد وضعیت متقاضیان مجدد، بحث و مذاکره کنیم؛ زیرا مخالفین همواره می‌توانند باعث برهم‌خوردن برنامه‌ریزیهای از پیش تعیین شده شوند (Conrad, 1994). برای مثال، برنامه بنیاد راکی‌فلر برای حمایت از زنان سرپرست خانوار توانایی پرداخت سریع کمک هزینه به تمامی متقاضیان را ندارد. در نتیجه، تخصیص تصادفی به عنوان راهی قابل‌قبول به لحاظ اخلاقی برای توزیع خدمات در میان زنان واجد شرایط پذیرفته شد (Boruch, 1997). اگرچه برخی مدیران برنامه محلی مخالف بوده، و به جای محدود کردن تعداد زنان دریافت‌کننده منابع خود را به تعداد بسیار زیادی زن تخصیص داده اما در عوض به هر یک، مقدار اندکی خدمات ارائه نمودند. در نهایت، اگر بخش بزرگی از متقاضیان رد شده احتمال تقاضای مجدد و پذیرفته شدن در مداخله را دارند، آزمایش تصادفی چندان توجیه‌پذیر نخواهد بود. اگر احتمال دارد سهم متقاضیان موفق کوچک باشد، روشهایی که در فصل ۱۰ برای مقابله با مشکلات اجرای مداخله به آنها اشاره شد، می‌تواند مفید باشد.

زمانی که یک نوآوری را نمی‌توان در زمان واحد روی تمامی واحدها اجرا کرد

اغلب به لحاظ فیزیکی یا مالی امکان ارائه یک نوآروی به تمامی واحد ها در یک زمان واحد وجود ندارد. چنین موقعیتهایی در حوزه آموزش و پرورش زمانی رخ می‌دهد که یک طرح درس جدید به آرامی یا به تدریج تغییر داده می‌شود؛ زیرا ابزارهای آموزشی جدید به آرامی و همزمان با معرفی کامپیوتر و یا روشهای جدید آموزش ضمن خدمت معلمین در سطوح مختلف سیستم اجرایی می‌شوند. در چنین موقعیتهایی، آزمایش می‌تواند

نوآوری را در چند مرحله ارائه نماید؛ به این صورت که برخی از افراد به طور تصادفی نوآروی را پیش از دیگران دریافت نمایند. این خود می‌تواند بوجودآورنده حالتی شبیه گروه کنترل و گروه آزمایش داشته باشد؛ البته تا زمانی که نوبت به دریافت نوآروی توسط گروه کنترل فرا می‌رسد. حتی بهتر خواهد بود اگر اینکار را بتوان با استفاده از طرح تکرارهای جابجا شوند که برای طرحهای شبه‌آزمایشی توضیح داده شد، انجام دهیم؛ البته با تکرارهایی که حالا دیگر به طور تصادفی تخصیص داده می‌شوند.

زمانی که واحدهای آزمایشی را می‌توان به طور موقت ایزوله کرد: طرح نمونه‌های هم‌ارز از نظر زمانی[۴۶۶]

اگرچه ما عموماً به تخصیص تصادفی افراد، مدارس، جوامع یا شهرها به شرایط مختلف آزمایشی و کنترل فکر می‌کنیم، همچنین می‌توانیم زمانها را به طور تصادفی به شرایط تخصصیص دهیم (Hahn, 1984). کمپبل و استنلی (Campbell & Stanley, 1963) این کار را «طرح نمونه‌های هم‌ارز از نظر زمانی» می‌نامند، تا بر این مسأله تاکید کرده باشند که تصادفی‌سازی بازه‌های زمانی‌ای که در آنها مداخله وجود داشته و نداشته (حاضر یا غایب بوده است)، را برابر می‌نماید. ادگینگتون (Edington, 1987) مثالهای متعددی از طرحهای با یک شرکت کننده منفرد ارائه می‌دهد، که در آنها، مداخلهها به طور تصادفی، در بازه‌هایی از زمان ارایه‌شده، و حذف می‌شدند. از آن جمله می‌توان به مقایسه سه دارو برای حمله خواب یا نارکولپسی، درمان با دارونما برای یکی از اختلالات دستگاه گوارش، و اثرات رنگ کردن مصنوعی غذا با دارونما بر رفتار کودکان بیش‌فعال، اشاره داشت. اثر باید دوره کوتاهی داشته باشد، بطوریکه اندازه اثر آن با حذف مداخله کاهش بیابد؛ و اثر باید در مواجهه مکرر، مداوماً به پاسخ ادامه دهد، به طوریکه هنگام اجرای مجدد مداخله، افزایش پیدا کند.

اما این اصل تنها مختص طرحهای با تنها یک شرکت‌کننده نیست. می‌توان از آن هنگامی که چرخش‌های طبیعی در گروهها وجود دارد، و هر گروه در زمان موردنظر از دیگر گروهها ایزوله شده است، مورد استفاده قرار بگیرد. بنابراین، زمانی که ۲۴ گروه از افراد برای اقامتهای نوبتی ۲ هفته‌ای در یک مرکز مشاوره روستایی اعلام آمادگی کردند، ماز (Mase, 1971) به طور تصادفی آن گروهها را به یکی از دو نوع از آموزش حساسیت تخصیص داد؛ یک آموزش برای دوازده گروه. در این مثال، خلق همزمان شرایط مداخله و کنترل می‌توانست منجر به انتشار مداخله و یا دیگر تهدیدهای عکس‌العملی روایی شود؛ اما طرح نمونه‌های هم‌ارز زمانی از این قبیل مشکلات اجتناب می‌کند. گرچه در نظر داشته باشید که، شرکت‌کنندگان الان در نمونه‌های زمانی به همان شیوه‌های لانه‌گزینی شدند که ممکن بود در یک مجموعه مانند مدرسه، یا یک منطقه جغرافیایی لانه‌گزینی می‌شدند. این نکته را باید در تحلیلها در نظر گرفت.

[466] Equivalent-time-samples

زمانی که واحدهای آزمایشی از نظر مکانی مجزا هستند، و یا سطح مراودات بینافردی پایین است

زمانی که واحدها از نظر جغرافیایی پراکنده هستند، و روابط حداقلی با یکدیگر دارند، و یا زمانی که می‌توان این شرایط را برای آنها ایجاد کرد، می‌توان این واحدها را به صورت تصادفی تخصیص داد. این اغلب در سازمانهایی اتفاق می‌افتد که تعداد زیادی شعبه دارند؛ برای مثال، سوپر مارکتها، واحدهای نظامی، دپارتمانهای دانشگاهی، مدارس درون محدوده‌های آموزشی، اتاقهای عمومی بستری بیماران در بیمارستانها، واحدهای اقامتی مذهبی، شعب کلوپهای سلامت در شهرهای بزرگ، و از این دست. اگرچه ایزوله کردن مکانی نمی‌تواند تضمین‌کننده ارتباط حداقلی باشد؛ در نتیجه، باید مراقب بود که این حالت تفکیک و بی‌ارتباطی حتماً به درستی رعایت شود.

برای مثال، آزمایشی در پرو اثرات ارایه خدمات مامایی و برنامه‌ریزی خانواده به مراجعین ۴۲ کلینیک که از نظر جغرافیایی پراکنده بودند، را مورد بررسی قرار داد (Population Council, 1986). کلینیکها به طور تصادفی به شرایط دریافت یک، دو، یا چهار ویزیت دکتر در هر ماه تخصیص داده شدند. جدایی جغرافیایی کلینیکها به این معنا بود که زنان در طول زمان به یک کلینیک سر می‌زدند، و در نتیجه شانس انتشار مداخله بسیار اندک بود. حتی اگر انتشاری امکان بروز داشت (مثلاً اگر زنی به طور معمول به دو کلینیک نزدیک به یکدیگر مراجعه می‌کرد)، محققین می‌توانستند کلینیکها را بر اساس محدوده‌های جغرافیایی مسدود (بلاک) کرده و محدوده‌های کلینیکی را (به جای خود کلینیک) به طور تصادفی تخصیص دهند. پرنگ (Perng, 1985) افراد را به شش روش مختلف که خدمات درآمد درونی برای دریافت بازده مالیات لحاظ می‌کرد، تخصیص داد. اغلب افراد از نظر جغرافیایی مجزا بودند. اما حتی اگر از نظر جغرافیایی به یکدیگر نزدیک بودند، با در نظر گرفتن اینکه از نظر قانونی اطلاعات بازده مالیات خصوصی و سرّی محسوب شده، و افراد به طور کلی از افشای اطلاعات مالیاتی خود اجتناب می‌کنند، غیرمحتمل بود که ارتباطاتی میان افراد واقع در شرایط مختلف آزمایش و کنترل رخ دهد. این آزمایشها یک نقطه‌قوت دیگر نیز داشت. هر دو آنها از ظاهر طبیعی درمانها برای تصادفی‌سازی مداخلهها (بدون اینکه جلب توجه کرده باشند[۴۶۷]) بهره برده بودند. بیماران انتظار این را داشتند که پزشکها به کلینیکها سر بزنند، و کمتر احتمال داشت که بیماران متوجه تغییرات جزیی در تعداد دفعات انجام این ملاقاتها شوند. افرادی که نامه‌های مالیات را از خدمات درآمد درونی دریافت می‌کردند، به ندرت با رویه این مؤسسه آشنا بودند تا بدانند که تغییری در رویه‌های معمول رخ داده. جلب توجه نکردن هدف ارزشمندی است، که محققین باید برای حصول آن تلاش کنند، به غیر از مواردی که مداخله عمداً طوری طراحی‌شده که با انتظارات پاسخ‌دهندگان متفاوت باشد.

[467] Unobtrusive

زمانی که تغییرات الزامیست و راه حلها ناشناخته هستند

برخی مواقع، تمامی گروههای دخیل توافق می‌کنند که یک شرایط نامطلوب نیازمند تغییر است، اما علیرغم طرفداری گروههای ذینفع از برخی گزینه‌ها، مشخص نیست چه تغییری باید انجام شود. اگر شرایط اجرایی، سیاسی، و اقتصادی اجازه دهد، امتحان کردن گزینه‌های مختلف تغییرات در یک آزمایش رسمی، احتمال بیشتری برای جلب موافقت از سوی همه طرفها را خواهد داشت. سوء استفاده از همسر یکی از بزه‌کاریهای جدی است که می‌تواند به قتل همسر منجر شده و بنابراین ماموران پلیسی که به محل جنایت فراخوانده می‌شوند، باید برخی اقدامات را انجام دهند. اما در اینکه اقدام موردنظر چه باید باشد توافق وجود ندارد؛ آیا باید میان زن و شوهر در همان جا اقدامات مشاوره‌ای انجام شود، آیا باید فرد آسیب‌رسان را وادار کرد که مکان را برای ۸ ساعت ترک کند، و یا باید فرد آسیب‌رسان را دستگیر کرد. یکی از مدیران که نظر مساعدی نسبت به روشهای آزمایشی برای یافتن راه حل مشکل داشت، اجازه داد تا یک آزمایش تصادفی انجام شده و بررسی شود کدامیک از این راه حلها نتیجه بهتری خواهند داشت. به همین طریق، در جایی دیگر آزمایشی تصادفی برای درمان بیماران با مشکلات روانی عدیده از طریق روش مراقبتهای استاندارد، یا روشی کاملاً متفاوت می توانست انجام داده شود، زیرا تمامی طرفها پذیرفته بودند که اطمینان ندارند کدامیک از درمانها برای این بیماران عملکرد بهتری خواهد داشت (Test & Burke, 1985).

اگرچه اینگونه مطالعات با هدف بررسی مداخلههای متفاوت می‌توانند نتایج ارزشمندی دربرداشته باشند، هر واریته ممکن است دقیقاً اهدافی در راستای اهداف برنامه اصلی نداشته باشد. در آزمایش سوءاستفاده از همسر مینیاپولیس، این عدم‌توافق بالقوه مسأله‌ساز نبود، چون بیشتر طرفها بر این موضوع توافق داشتند که هدف مطلوب در نهایت کاهش در تکرار خشونت پس از انجام مداخله است. اگرچه، عدم توافق بیشتر زمانی می‌تواند رخ دهد که شرکت‌کنندگان به پروژه‌هایی تخصیص داده شوند که مدیران، کارکنان، و سرمایه‌گذارانی متفاوت داشته باشند (در قیاس با پروژه‌هایی که در آنها تمامی واریته‌ها توسط افرادی یکسان اجرا می‌شوند). همچنین مدیران پروژه‌های مختلف همواره درباره مقیاسهایی که برای اندازه‌گیری مفاهیم موضوع تغییر بکارگرفته می‌شوند، توافق ندارند.

زمانی که یک گره قابل گشودن بوده، و ابهام در مورد نیاز قابل برطرف کردن باشد

تخصیص افراد به شرایط بر اساس نیاز و یا شواهد در مقایسه با تخصیص بر اساس تصادفی‌سازی اغلب راه و قائده جذابتری برای مدیران، کارکنان، و شرکت‌کنندگان است. چنین ملاحضاتی را میان توجیحات وارده برای طرح ناپیوستگی رگرسیون می‌توان مشاهده کرد. اگرچه نیاز یا شواهد برخی افراد اغلب مبهم است. در این موارد، ابهام را می‌توان از طریق تخصیص تصادفی افراد با نیازهای مبهم به شرایط (شاید در ترکیب با طرح ناپیوستگی رگرسیونی) برطرف نمود. به همین طریق، لیپسی و همکاران (Lipsey et al., 1981) تخصیص تصادفی

را به عنوان راهی برای برطرف کردن ابهام در ارزیابی یک برنامه هدایت نوجوانان بزهکار مورد استفاده قرار دادند. در یک طرح شبه‌آزمایشی، ماموران پلیس بیشترین تلاش خود را مصروف اینکار کردند که تشخیص دهند آیا نوجوان دستگیرشده را باید (۱) مورد خدمات مشاوره قرار داد و بعد وی را آزاد کرد، (۲) به دوره‌های بازآموزی و کارورزی فرستاد، (۳) و یا اینکه به پروژه های فشرده‌تر و جدی‌تر خدمات اجتماعی که همزمان مشاوره، آموزشهای درمانی، بازپروری و خدمات ترک اعتیاد را در برمی‌گیرند فرستاد. اگرچه، هنگامی که ماموران مطمئن نبودند کدام تخصیص بیشتر از همه مورد نیاز است، و به این قضاوت می‌رسیدند که هرکدام از مشاوره و آزادی و یا بازآموزی و کارورزی می‌تواند مفید باشد، نوجوانان را به صورت تصادفی به یکی از این دو شرایط تخصیص می‌دادند.

در چنین آزمایشات گره‌گشایی، تعمیم دادن محدود به افرادی می‌شود، که در محدوده نیاز مبهم نمره می‌گیرند، گروهی که کمترین اطلاعات و دانش را در مورد درمان اثربخش برای آنها در اختیار داریم. اگرچه، اگر یک سازمان در زمینه درمان بهترین، بدترین، و یا کل طیف شرکت‌کنندگان متخصص باشد، کارکنان آن به خوبی می‌دانند که ارزیابی عملکرد آنها برای شرکت‌کنندگان مبهم، از آنچه آنها در واقع انجام می‌دهند، تأثیر نمی‌پذیرد. خوشبختانه ممکن است بتوان یک آزمایش گره‌گشا را به برخی شبه‌آزمایشهای قابل تفسیر ربط داد؛ مانند آنچه که لیپسی و همکارانش (۱۹۸۱) برای دستیابی به این اهداف انجام دادند.

هنگامی که برخی افراد هیچ‌کدام از گزینه‌ها را ترجیح نمی‌دهند

حتی اگر اخلاقیات و قوائد عمومی این الزام را بوجود بیاورند که افراد باید اجازه داشته باشند خودشان گزینه‌ای که دریافت خواهند کرد را انتخاب کنند، افرادی که هیچکدام از گزینه‌ها به طور ویژه برایشان ارجح نیست را می‌توان بر مبنای شانس، تخصیص داد. برای مثال، والین و باوم (Valins & baum, 1973) می‌خواستند برخی از اثرات محیط فیزیکی بر دانشجویان سال اول که به یکی از دو نوع اتاق وارد می‌شدند، را بررسی کنند؛ این دو اتاق از نظر تعداد افردی که فرد ممکن بود در طول روز ملاقات کند با یکدیگر تفاوت داشتند. محققین مطالعه را محدود به ۳۰٪ کسانی کردند که برایشان تفاوتی نداشت در کدام نوع از این اتاقها اقامت کنند. مسئولان خوابگاهها این ۳۰٪ را به صورت اتفاقی و کاتوره‌ای به اتاقها تخصیص دادند؛ اما احتمالاً بهتر بود اینکار را به صورت تصادفی انجام می‌دادند. بدون تردید، محدود کردن آزمایش به افرادی که ترجیح مشخصی نسبت به اتاقها نداشتند، تعمیم نتایج به افرادی غیر از این افراد را با محدودیت مواجه می‌کرد. اگر افراد دارای ترجیح و بی‌تفاوت، همگی موردنظر محقق باشند، آزمایش تصادفی بر روی افراد بدون ترجیح را می‌توان در کنار یک شبه‌آزمایش برروی افراد دارای ترجیح انجام داد. سپس نتایج دو مطالعه را مقایسه نمود؛ بطوریکه ضعف یک مطالعه، قوت مطالعه دیگر باشد. اگر نتایج شبیه یکدیگر باشد، آنگاه می‌توان یک استنباط کلی بدست آورد.

هنگامی که می‌توانید سازمان خودتان را خلق کنید

تخصیص تصادفی جزءی پذیرفته‌شده از فرهنگ سازمانی آزمایشهای لابراتواری است. اما اغلب آزمایشهای میدانی در فرهنگهای سازمانی‌ای انجام می‌شوند که در آن، تصادفی‌سازی عنصری بیگانه است. با این حال برخی اوقات محققین می‌توانند سازمانهایی برای خود خلق کنند که در آنها بتوانند تصادفی‌سازی را به رویه‌ای معمول تبدیل کنند. برای مثال، دپارتمانهای روانشناسی دانشگاهها اغلب مراکز خدمات روانشناسی برای تسهیل آموزش دانش‌آموزان فارغ التحصیل رشته روانشناسی کلینیکی دایر می‌کنند؛ اینکار به اعضای دپارتمان روانشناسی اجازه می‌دهد تا کنترل آزمایشی بیشتری از آنچه در کلینیکهای روانشناسی معمول است، را بر کلینیکها اعمال کنند (Beutler & Crago, 1991). در چنین مراکزی، محققین می‌توانند کنترل بهتری نه تنها بر تصادفی‌سازی بلکه بر مؤلفه‌هایی مانند استانداردسازی مداخله، اندازه‌گیریها و انتخاب نمونه داشته باشند. مؤسسات تحقیقاتی مستقل و مراکز متمرکز بر مشکلات ویژه، مکرراً امکان برقراری سطوح مشابهی از کنترل وسیع را فراهم می‌آورند. برای مثال، خط تلفن کمک به سیگاریها در کالیفرنیا جلسات رایگان ترک سیگار در ایالتهایی که در آنها می‌شد با خط تلفن مذکور تماس گرفت را ارائه می‌کردند (Zhu, 1999). تصادفی‌سازی تماس‌گیرندگان به گروه‌های درمان و کنترل به لحاظ اقتصادی توجیه‌پذیر نبود. تمامی تماس‌گیرندگان یکی ایمیل درمانی را دریافت می‌کردند که در آن دستورالعملی برای تماس مجدد، زمانی که آمادگی شروع درمان را داشته باشند، آورده شده بود. سپس افرادی که مجدداً تماس نمی‌گرفتند به دو گروه تخصیص داده می‌شدند: گروهی که در مورد آنها اقدام دیگری انجام نمی‌گرفت، و گروهی که از سوی کارکنان درمان با آنها تماس گرفته می‌شد تا درمان را آغاز کنند. از ابتدا این رویه می‌توانست برای تقسیم تصادفی غیرپاسخ‌دهندگان به دو گروه در هر شبه‌آزمایش، برای تقسیم گروه مداخله به دو گروه مداخله و کنترل مورد استفاده قرار بگیرد. مثلاً برای کسانی که رواندرمانی درخواست کرده بودند، اما نتوانسته بودند در جلسات حاضر شوند، یا برای کسانی که دستورالعمل را دریافت کرده بودند، اما نتوانسته بودند به آن عمل کنند، و یا آنهایی که برای یک دوره آموزش ضمن‌خدمت پذیرفته‌شده بودند، اما نتوانسته بودند در آن حاضر شوند، و از این دست. در نهایت، محققین برخی اوقات می‌توانند سازمانهایی را تنها برای کنترل تصادفی‌سازی پایه‌گذاری کنند؛ مانند آنچه که اغلب در آزمایشهای چندمکانی پزشکی اتفاق می‌افتد؛ که در آنها یک ستادمشترک مرکزی که توسط محقق کنترل می‌شود، ساخته می شود تا تصادفی‌سازی را انجام دهد. پروژه مشارکتی افسردگی متعلق به موسسه ملی سلامت روان از این روش ستاد مرکزی برای کنترل تصادفی‌سازی بهره برد (Collins & Elkin, 1985).

هنگامی که محقق بر واحدهای آزمایشی کنترل دارد

امکان پایه‌گذاری سازمان و یا ایجاد ستاد مرکزی تصادفی‌سازی بسیار نادر است. اغلب محققین میدانی مهمان موسسات تحقیقاتی دیگر هستند، و بسیاری از امکانات کنترلی خود را از میزبانان قدرتمند خود دریافت می‌کنند. نمونه‌ای از این حالت، زمانی اتفاق افتاد که راه‌حلهایی برای مشکل ساعات اوج مصرف برق بوسیله شرکتهای آب و برق مورد ارزیابی قرار می‌گرفت. مصرف برق بسته به زمان روز تفاوت می‌کند، و شرکتهای آب و برق باید ظرفیت کافی برای پاسخ‌گویی به تقاضای زمان اوج مصرف داشته باشند، حتی اگر این ظرفیت در دیگر زمانهای روز به طور عمده بلا استفاده باقی می‌ماند. ایجاد این ظرفیت پرهزینه است. در نتیجه شرکتهای آب و برق می‌خواستند بدانند آیا دریافت بهای بیشتر برای برق در زمان اوج مصرف می‌تواند تقاضا را کاهش داده و در نتیجه نیاز برای ساختن ظرفیت بیشتر را برطرف سازد؟ محققین قادر بودند خانوارها را به طور تصادفی به شرایط تقاضای زمان اوج مصرف در مقابل شرایط نرخهای استاندارد تخصیص دهند؛ زیرا میزبانهای آنها (یعنی شرکتهای برق) بر عرضه خدمات به خانوارهای مورد نظر کنترل کامل داشتند، و همچنین نسبت به نتایج آزمایش در پاسخ به سؤال مورد نظر علاقمند بودند.

تصادفی‌سازی همچنین زمانی بیشتر امکان‌پذیر است، که پایه‌گذاران اصلی بر آن پافشاری می‌کنند. برای مثال، موسسه ملی سوء مصرف مواد مخدر و موسسه ملی سوء مصرف الکل بودجه‌هایی را در قالب گرنتهای تحقیقاتی به خدمات نوآورانه‌ای که با استفاده از طرحهای تحقیقاتی آزمایشی مستحکم مورد ارزیابی قرار گرفته بودند، تخصیص می‌دادند (Coyle, Boruch & Turner, 1991). اگرچه خصوصاً زمانی که سرمایه گذاران و گرنت گیرندگان رابطه بلند مدتی دارند، وجود شیوه های کنترل هزینه های مالی از سوی سرمایه گذاران می تواند به تنشهایی منتهی شود. به عنوان نمونه، لم و همکارانش (Lam et al., 1994) متوجه نوعی «وابستگی متقابل تنش زا»[۴۶۸] میان دانشگاه ییل و شهر نیوهون (که ییل در آن واقع شده است) شدند. این وابستگی به دلیل این واقعیت بود که محققین ییل مکرراً انواع خدمات اجتماعی را به شهر ارائه می‌کردند؛ خدماتی که امکان ارائه آنها وجود نداشت، مگر همراه با زنجیره پژوهشی متصل به آن. در مجموع، انگیزه میزبان و شرکت‌کنندگان برای داوطلب‌شدن برای آزمایش فردا تا حد زیادی به نحو برخورد با آنها در آزمایش امروز وابسته است.

هنگامی که انتظار انجام قرعه‌کشی وجود دارد

برخی مواقع از قرعه‌کشیها به عنوان ابزاری پذیرفته‌شده از نظر اجتماعی برای توزیع منابع استفاده می شود. از جمله این موارد می‌توان به قرعه‌کشی بکارگرفته‌شده برای تخصیص دانش‌آموزان دختر به خوابگاهها در دانشگاه استنفورد (Siegel & Seigel, 1957)، قرعه‌کشی برای انتخاب میان متقاضیان ورود به یک مدرسه مگنت، و قرعه

[468] Contentious codependence

کشی قانون ۱۹۷۰ در سال در آمریکا (Notz, Staw, & Cook, 1971) اشاره داشت. در این موارد، انگیزه تصادفی‌سازی انجام تحقیق نبود، بلکه در عوض سرمایه‌گذاری بر این ادراک و نگرش که تصادفی‌سازی راهی بدون‌سوگیری برای توزیع منابع است. این قبیل کاربردهای اجتماعی تصادفی‌سازی نوعی آزمایش تصادفی طبیعی می‌سازد که محققین می‌توانند آن از آن بهره‌برداری کنند. متاسفانه، قرعه‌کشیهای رسمی اجتماعی به طور مکرر اتفاق نمی‌افتند در نتیجه، غالب وقتها نمی‌توان به آنها به عنوان ابزاری برای خلق گروههای هم‌ارز بصورت احتمال-پایه تکیه کرد.

هنگامی که تخصیص تصادفی توجیه اقتصادی ندارد و یا مطلوب نیست

حتی زمانی که تمایل به بررسی میزان اثربخشی یک مداخله وجود دارد نیز، ممکن است شرایطی بوجود بیاید که امکان بهره‌گیری از آزمایشهای تصادفی را محدود نماید. از جمله این شرایط می‌توان به چند مورد اشاره نمود. اول، آزمایشهای تصادفی ممکن است زمانی که به دنبال پاسخ سریع به سؤال هستیم، مطلوب نباشند. عموماً سالهای متمادی میان آغاز تعریف یک آزمایش میدانی بزرگ، و به دست‌آمدن نتایج نهایی آن فاصله وجود دارد- علی‌الخصوص اگر درمان نیازمند زمان باشد (مانند آنچه در رواندرمانی‌های طولانی اتفاق می‌افتد)، و اگر نتایج میان‌مدت و بلندمدت موردنظر محقق باشد (مانند درآمد در طول عمر فرد). برای مثال، در آزمایش مالیات بر درآمد منفی نیوجرسی، «چهار سال فاز اولیه میان ۴۴ ماه برنامه‌ریزی و طراحی و ۱۶ ماه جمع‌آوری داده‌ها گذشت»- یعنی چیزی حدود ۸ سال طول کشید (Haveman, 1987). در نتیجه، اگر اطلاعات به سرعت موردنیاز بود، گزینه‌های جایگزین آزمایش تصادفی، احتمالاً عملکرد بهتری می‌داشتند. مثلاً دپارتمان ارزیابی برنامه و روش در اداره کل حسابداری ایالات متحده مکرراً با سوالهای متعددی از سوی قانونگذارانی مواجه است که درصدد تصمیم‌گیریهای سریع هستند. برخی از این سؤالات دربرگیرنده اثرات برنامه‌ها یا سیاستگذاری هاست. تأخیر چند ساله در پاسخ‌گویی می‌تواند تصمیم‌گیریها را بیش از اندازه طولانی نماید؛ بطوریکه تصمیم موردنظر دیگر موضوعیت نداشته، و سؤال‌کننده دیگر در مسند کار نباشد. در نتیجه، این دپارتمان ندرتاً از آزمایشهای تصادفی استفاده کرده، و در عوض بر ترکیبی از شبه‌آزمایشها، پیمایشها و مطالعات مروری در مورد اثرات سیاستهای مربوطه استفاده می‌کند (Chan & Tumin, 1997; Datta, 1997; Droitcour, 1997). چنین رویه‌هایی برای استنباط در مورد علل یک پدیده، ممکن است ضعیف‌تر از یک آزمایش تصادفی باشند؛ زیرا حتی زمانی که ادبیات دربرگیرنده نتایج مطالعات آزمایشی است نیز، به ندرت اتفاق می‌افتد که مطالعه‌ای آزمایشی عیناً در مورد سؤال تحقیق موردنظر انجام شده‌باشد. اما روشهای GAO غالباً همواره از صحت قابل قبولی برخوردارند و نتایج آنها سریعتر از نتایج آزمایشهای تصادفی آماده می‌شود.

دوم، آزمایشهای تصادفی پاسخی دقیق به این سؤال که آیا مداخله موردنظر عمل کرده است یا نه می‌دهند (Cronbach et al. 1980). اما بسیاری از مطالعات به دقت بسیار زیاد نیازی ندارند. برای مثال، زمانی که اطلاعات پیش‌زمینه‌ای با کیفیت بالا در مورد مداخله مورد نظر وجود دارد، انجام مرور ادبیات موجود می‌تواند گزینه بهتر و کم هزینه‌تری باشد (در مقایسه با انجام یک آزمایش تصادفی جدید). در زمانی که یک سؤال علّی در رتبه ثانویه اهمیت در قیاس با یک سوال غیرعلّی قرار داشته باشد (مانند اینکه آیا خدمات همانطور که مورد نظر بوده ارائه شده اند، در قیاس با پایش رویه های ارائه خدمات). زمانی که یک اثر آنچنان بزرگ و غیرمعمول است که تردیدی باقی نخواهد گذاشت که در اثر مداخله بوجود آمده است؛ مانند حالتی که در مورد اثر بزرگ غربالگری PKU و عقب‌ماندگی ناشی از آن در کودکان رخ داد. در این حالت، انجام یک آزمایش تصادفی دیگر بلاموضوع است.

سوم، آزمایشهای تصادفی به ندرت می‌توانند برای پاسخ به انواع خاصی از سؤالات بکار گرفته شوند. امکان ندارد بتوان افراد را به طور تصادفی به متغیرهایی تخصیص داد که امکان دستکاری آنها وجود ندارد (مانند سن، نژاد). همچنین نمی‌توان رویدادهایی را که در گذشته رخ داده‌اند را دستکاری کرد (مانند اثرات رکود اقتصادی ۱۹۳۰ در ایالات متحده). اینکار غیراخلاقی خواهد بود که افراد را به طور تصادفی به رویدادهای قابل دستکاری که موجب آسیبهای فراوان می‌شوند (مانند سیگار کشیدن یا داشتن آسیب جدی نخاعی) تخصیص دهیم.

چهارم، قبل از انجام یک آزمایش مقدار قابل‌توجهی کار مفهومی و کاربردی باید انجام شود. مرکز قضایی فدرال (۱۹۸۱) توصیه می‌کند قبل از آنکه یک آزمایش انجام شود، باید محرز شود که (۱) شرایط فعلی نیازمند ارتقاء و رشد است، (۲) ارتقاء پیشنهاد شده عواقب نامشخصی دارد، (۳) تنها یک آزمایش می‌تواند اطلاعات لازم برای تغییر سیاستها را فراهم آورد، (۴) نتایج آزمایش برای تغییر کاربردها و سیاستها بکارگرفته می‌شود، و (۵) به حقوق افراد در خلال آزمایش احترام گذاشته می‌شود. به همین طریق، مدل پنج مرحله‌ای آزمون روشهای کنترل سرطان پیشنهاد می‌کند قبل از انجام یک آزمایش تصادفی، (۱) متون علمی موجود باید مشخص و با یکدیگر ترکیب شود تا ببینیم آیا می توان فرضیاتی قابل‌پشتیبانی و قابل‌آزمون تولید کرد؟ (۲) باید آزمونهای پایلوت انجام شود تا توجیه‌پذیری اقتصادی یا قابل‌پذیرش‌بودن یک درمان مورد بررسی قرار گیرد، (۳) باید مطالعاتی در رابطه با ارزیابی مشارکت و پذیرش مداخله در جامعه موردنظر انجام شود، (۴) اشکال جمع‌آوری داده باید ساخته‌شده و اعتبارسنجی شوند؛ و (۵) مطالعات کنترل‌شده بوسیله شبه‌آزمایشها را باید برای ارائه شواهد اولیه در زمینه اثرات مداخله بکار گرفت (Greenwald & Cullen, 1984). آزمایش کردن ناپخته تنها هزینه‌کرد مقادیر فراوانی از منابع است؛ و البته می‌تواند اثر مداخلههای بالقوه اثربخشی را که بطور نادرست مورد آزمون قرار گرفته‌اند را بی‌اهمیت جلوه دهد.

بحث و نتیجه گیری

آزمایش تصادفی اغلب روشی ارجح برای بدست‌آوردن تخمین بدون‌سوگیری از اثرات یک مداخله یا مداخله است. این آزمایشها نسبت به دیگر روشها فرضیات کمتری دارند، اعتبار آن فرضیات را معمولاً راحتتر می‌توان بر اساس داده‌ها سنجید، و در آنها، دانش پیشین کمتری درباره مواردی همچون فرایندهای انتخاب و مشخصه‌های واحدها (که به طور معمول در شبه‌آزمایشها، مدلسازی علّی و مدلهای سوگیری انتخاب لازم است) مورد نیاز است. با در نظر گرفتن تمامی این مزایا و قوتها، محقق ممکن است به آسانی مشکلات اجرایی که می‌تواند در جریان انجام اینگونه آزمایشها رخ دهد را نادیده بگیرد.

یکی از مشکلات اجرایی، توجیه‌پذیری اقتصادی و مطلوبیت این روشها در زمینه برخی موضوعات است. برخی دستکاریهای آزمایشی اخلاقی نیستند و بنابراین قابل تصادفی‌سازی نیستند؛ مانند اینکه پزشکی تصمیم بگیرد نوع خاصی از بیماران باید درمان مشخصی را دریافت کنند. یا اینکه یک مداخله آزمایشی آنچنان اثرات منفی و مثبتی ایجاد می‌کند، که مطالعه آنها روی افراد اخلاقی نباشد. در طول زمان، این قابل قبول نیست که سالها صبر کنیم تا یک آزمایش به درستی طراحی شده و اجرایی شده بتواند انجام شود. همینطور در مواقعی مشکلات حقوقی پیش می‌آیند؛ نه تنها چون زیرپاگذاشتن اخلاقیات می‌تواند مشکلات حقوقی ایجاد کند، بلکه چون برخی آزمایشها تحت قوانین خاصی انجام می‌شوند. برای مثال، زمانی که آزمایش برای ارزیابی آزمایشی یک برنامه انجام می‌شود، یا زمانی که شرکت‌کنندگان به طور مستقیم تحت محافظت حقوقی هستند (مانند زندانیان)، و یا زمانی که خود سیستم حقوقی موضوع مطالعه است.

دومین مشکل اجرایی این است که ممکن است تعداد کافی افراد (واحدها) واجد شرایط که مایل به انجام مداخله هستند (در صورتی که به طور رندم به شرایط مداخله تخصیص داده شوند)، وجود نداشته باشد. بسیاری از آزمایشها به همین دلیل با شکست مواجه می‌شوند. به طور مکرر، علی‌الخصوص محققینی که هیچگاه پیش از این آزمایشهای میدانی بزرگ انجام نداده‌اند، تعداد افراد واجد شرایط و سهولت جای‌دهی به این افراد را بیش از واقع تخمین می‌زنند. حتی زمانی که این افراد جای داده می‌شوند، آنها اغلب از شرکت در آزمایش سرباز می‌زنند. در بدترین حالت، آزمایش به دلیل نبود شرکت‌کننده عقیم مانده، و از بین می‌رود.

سومین مشکل کاربردی این است که رویه تصادفی‌سازی همواره به طور درست طراحی و اجرا نمی‌شود. برخی مواقع این مشکل به این دلیل ایجاد می‌شود که محقق به درستی تخصیص تصادفی را درک نکرده است و رویه‌ای مانند تخصیص کاتوره‌ای و نامنظم را به جای رویه تخصیص تصادفی اجرا می‌کند. و یا اینکه محقق تعدیلهای دیگری را پس از تخصیص تصادفی انجام می‌دهد که باعث می‌شود گروهها قبل از شروع آزمایش با یکدیگر متفاوت باشند؛ و در تمام طول مدت آزمایش فرض را بر این بگذارند که این رویه‌ها تصادفی بوده در حالی که در واقع نبوده است. برخی دیگر از مواقع رویه‌های تخصیص تصادفی به درستی طراحی می‌شوند اما

محقق نمی‌تواند به درستی آن را اجرا کند، و یا بر اجرای آن نظارت کند، در نتیجه تخصیص به درستی انجام نمی شود. زمانی که یکی از این دو حالت (برنامه‌ریزی نادرست یا اجرای ناکامل) اتفاق می‌افتد، مزیتهای تخصیص تصادفی همگی از دست می‌رود.

چهارمین مشکل کاربردی این است که مداخله تخصیص یافته، دقیقاً همان که فرد دریافت کرده نیست. ممکن است شرکت‌کنندگان مداخله مورد نظر محقق را به طور کامل دریافت نکنند و یا اینکه اصلاً آن را دریافت نکنند؛ مانند زمانی که بیماران تخصیص پیدا کرده به درمان دارویی، داروی مورد نظر را دریافت نمی‌کنند، و یا به طور ناقص دریافت می‌کنند. بیماران ممکن است به دیگر شرایط آزمایش سرریز[۴۶۹] شوند؛ مانند شرکت‌کنندگان گروه کنترل، که مجدداً برای درمان تقاضا می‌دهند، و پذیرفته می‌شوند. انتشار مداخله می‌تواند از طریق تعاملات و گفتگوهای میان شرکت‌کنندگان گروههای مختلف در مورد مداخله اتفاق بیافتد. در اینجا نیز، شرکت‌کننده بخشی از شرایط هر دو گروه را دریافت کرده است. در تمامی این موارد، تفاوتهای ناشی از مداخله که مورد نظر محقق است، از بین می‌رود. در چنین حالتی، حتی اگر این استنباط که تخصیص به شرایط موجب نتایج شده است، همچنان روشن باشد، روایی سازه مداخله روشن نیست. از این رو، درصورتی که شناسایی تمایزات خالص ناشی از مداخله مطلوب محقق است، لازم است تا از این گونه نقایص در اجرای مداخله پیشگیری کرده، و یا آنها را در محاسبات لحاظ نمایند.

پنجمین مشکل ریزش است. تخصیص تصادفی تنها با هدف مساوی کردن گروهها قبل از اجرای مداخله انجام نمی‌شود. بلکه این هدف را نیز دنبال می‌کند که گروهها در پس‌آزمون از همه جنبه‌ها بغیر از تفاوتهای ناشی از مداخله با یکدیگر برابر باشند. میزان متفاوت ریزش در گروهها بعد از انجام تخصیص تصادفی می‌تواند این هدف دوم را عقیم بگذارد. اینگونه ریزشها اغلب در مطالعات میدانی رخ می‌دهند. در نتیجه، پیشگیری از ریزش، مقابله با ریزش، محاسبه ریزش و تحلیل داده با در نظر گرفتن ریزش همگی اقدامات مهمی هستند که در مطالعات با آزمایشهای تصادفی باید در نظر گرفته شوند.

[469] Crossover

فصل ۱۳

استنباط علّی تعمیم داده شده: روشهایی برای [جمع بندی] مطالعات متعدد

متعدد (multiple): داشتن، مشتمل بودن، یا مرتبط بودن به بیش از یک فرد، عنصر، قسمت یا جزء؛ تعداد زیاد و متنوع (manifold)

Meta یا met: [کلمه ای یونانی از ریشه متا، به معنی ورای چیزی، بعد از چیزی] ۱. الف: متأخر در زمان (metestrus)؛ ب. در مراحل بعدی رشد (metanephros)؛ ۲. قرار گرفته در پس یا پشت چیزی (metacarpus)؛
۳. الف. تغییر و تغییر شکل (metachromatic)؛ ب. جایگزینی (metagenesis)؛ ۴. الف. ماورا یا فراتر از چیزی بودن، جامعتر (meta linguistics)، ب. در مرحله بالاتر از نظر رشد (metazoan)

اسمیت و گلس (Smith & Glass, 1977) طی یک مطالعه ای نتایج ۳۷۵ پژوهش صورت گرفته در رابطه با ارزیابی اثربخشی رواندرمانی را به صورت کمّی خلاصه و گزارش کردند. ۳۷۵ مطالعه ذکر شده طی دوره زمانی بسیار طولانی، در مکانهای جغرافیایی متفاوت، با مراجعین متنوع، و با استفاده از مداخله های درمانی و مقیاسهای اندازه گیری نتایج متفاوتی صورت گرفته بودند. با وجود تمامی این تفاوتها و تنوعها، مراجعینی که مداخله رواندرمانی را دریافت کرده بودند نسبت به مراجعینی که دریافت نکرده بودند (گروه کنترل)، نتایج بهتری بدست آوردند. نتایج درمان مشابه بود، خواه برای درمانهای رفتاری و خواه غیررفتاری، با هر میزان سطح تجربه درمانگر، و یا هر میزان زمان سپری شده برای درمان. اثرات رواندرمانی (بجز در مواردی معدود) علی رغم انواع مختلف متغیرهای تعدیلگر در مطالعات مختلف، همچنان قابل مشاهده و تعمیم بود. این مطالعه نمونه ای از یک متاآنالیز است، ابزاری قدرتمند برای انجام استنباطهای علّی تعمیم یافته. این روش که طی دهه

های اخیر پدید آمده و ارتقاءداده شده است، یکی از روشهایی است که در این فصل برای جمعبندی و تعمیم نتایج تعداد زیادی از مطالعات، مورد بحث قرار خواهد گرفت.

تعمیم بر مبنای نتایج یک مطالعه در مقابل تعمیم بر اساس نتایج چندین مطالعه

در فصل قبل تعمیم های مبتنی بر نتایج یک مطالعه را مورد بررسی قرار دادیم. متاسفانه یک مطالعه به ندرت طیف ناهمگونی از افراد، موقعیتها، زمانها، مداخلهها، و مقیاسهای ارزیابی مناسب برای تعمیم را در بر می گیرد. همچنین به ندرت یک مطالعه روشهای متنوعی را مورد استفاده قرار می دهد. در مقابل، چند مطالعه عموماً در تمامی این جنبه ها دارای تنوع بیشتری هستند. این تنوع در عناصر مطالعه به محقق اجازه می دهد تا آزمونهای بهتری بر روی استنباطهای علّی تعمیم داده شده انجام دهد.

برنامه های پژوهشی طرح ریزی شده برای بررسی متمرکز یک موضوع توسط یک محقق و یا در یک آزمایشگاه، یکی دیگر از راههایی است که به وسیله آن می توان به تنوع لازم برای انجام تعمیم دست یافت. چنین برنامه هایی بررسی مستقیم متغیرهایی که می توانند بر استنباط علّی تعیم داده شده اثر بگذارند، را میسر می سازد. محقق می تواند به طور سیستماتیک متغیرهای مهم را از مطالعه ای به مطالعه دیگر تغییر داده، و به تدریج به درکی دقیقتر و اصلاح شده از موضوع مورد تعمیم، متغیرهایی که تعمیم را محدود کرده، و یا مانع انجام آن می شوند، و متغیرهای واسطه گری که می توانند اثر مورد نظر را توضیح دهند، دست یابد. روش دیگر که از آن طریق می توان به ناهمگونی مورد نیاز برای انجام تعمیم ها دست پیدا کرد، جمع بندی کردن و خلاصه کردن مطالعات متعدد انجام شده توسط محققین مختلف در خصوص یک موضوع مشترک است، علی الخصوص خلاصه های کمّی با استفاده از متاآنالیزها. این خلاصه ها (در مقایسه با مطالعات منفرد) از یک سو تخمینهای دقیقتری از اثر موردنظر بدست می دهند و از سوی دیگر این امکان را به محقق می دهند که تغییرات رابطه علّی را با در نظر گرفتن عناصر مختلف در تحقیق رصد نماید. این تحلیلها کمک می کند تا ماهیت روابط علّی، حدود و مرزهای آن، رفتارهای آن در محدوده مرزها و تفسیرهای محتمل برای آن را روشن نماییم. بعلاوه، پایه وسیعتری که این دانش بر مبنای آن بدست آمده، نسبت به پایه دانش قابل دستیابی از یک مطالعه منفرد اعتبار بیشتری دارد. این موضوع زمانی بیشتر مصداق دارد که یک مداخله همواره بهتر از یک مداخله جایگزین عمل می کند و یا در شرایط و موقعیتهای مختلف به طور مستحکم و مثبت عمل می نماید. در مقابل، مداخله هایی که بعضی اوقات بهتر از جایگزین عمل می کنند و بعضی مواقع ضعیفتر، و یا مداخله ای که بعضاً آسیب زا می شوند، ارزش کمتری دارند، علی الخصوص اگر مقتضیات و شرایطی که منجر به بدست آمدن نتایج منفی می شود ناشناخته بوده، و یا نتوان آنها را به سادگی تغییر داد. در مباحثی که در ادامه این فصل ارائه می شود جزئیاتی را درباره راههای مختلفی که از آن طریق می توان استنباط علّی تعمیم داده شده را تسهیل نمود، مطرح می نمائیم.

برنامه های چند-مطالعه ای پژوهشی

برنامه های چند-مطالعه ای پژوهش کنترل قابل توجهی بر روی جنبه هایی از استنباط علّی تعمیم داده شده که مدنظر محقق است (از یک به مطالعه دیگر) می دهد، بنابراین محقق می تواند سوالاتی که در هر برهه از زمان نیازمند پاسخ

عاجل است را دنبال کند. اگرچه، این برنامه ها به نسبت دیگر روشهایی که مورد بحث قرار خواهیم داد به زمان و سرمایه گذاری مالی قابل توجهی نیاز دارند. برخی اوقات این سرمایه گذاری ها در مقیاسی بسیار کوچکتر، در اختیار محققینی قرار می گیرد که از حمایتهای بیرون و درون– سازمانی به طور همزمان بهره مند می شوند و برنامه های تحقیقاتی را در آزمایشگاههای خود هدایت می کنند.

مدلهای فازبندی شده مطالعات به طور فزاینده تعمیم‌پذیر[470]

یکی از رویکردهای موجود نسبت به استنباطهای علّی تعمیم یافته، مدل فازبندی شده است که در چندین شکل مختلف مورد استفاده قرار می گیرد. در آمریکا صندوقهای سرمایه گذاری متعددی تامین هزینه های تحقیقات مرتبط با سلامت را بر عهده دارند. این سازمانها و صندوقهای سرمایه گذاری اغلب از تحقیقاتی حمایت می کنند که فاز به فاز پیش می روند؛ از تحقیقات اکتشافی شروع شده و تا آزمون نتایج روی انسان پیش می روند. جزئیات این فازها از یک سازمان به سازمان دیگر، و از زمانی به زمان دیگر متفاوت است. اما توصیف گرین والد و کولن (Greenwald & Cullen, 1984) از این فازها در مطالعات مرتبط با سرطان تا حد زیادی در میان دیگر مطالعات نیز عمومیت دارد. قبل از اینکه فازها شروع شوند، تحقیقات پایه ای و اولیه در مورد اثرات یک مداخله روی انواع خاصی از سلولهای سرطانی در لوله های آزمایش و تومورهای تولید شده (به طور مصنوعی) در حیوانات انجام می شود. اگر نتایج آزمایش روی حیوانات مورد درمان قابل توجه باشد، مرحله (فاز) اول تحقیقات یعنی انجام تحقیقات پایه ای و کاربردی برای تدوین فرضیاتی آزمون پذیر در مورد درمان مورد نظر آغاز می شود. در فاز دوم به سمت ایجاد روشی حرکت می کنیم که تضمین کننده وجود فرایندی با روایی و درستی کافی برای مطالعه اثرات آن درمان باشد. فاز سوم، مداخلات کنترل شده برای بررسی اثربخشی درمان سرطان مورد نظر انجام می شود؛ به این معنی که بررسی می شود آیا این درمان تحت شرایط ایده آل که احتمالاً با شرایط کاربرد آن متفاوت است، نیز اثربخش خواهد بود یا نه. فاز چهارم شامل مطالعات ارزیابی اثربخشی با جمعیت تعریف شده است، برای اینکه ببینیم آیا درمان موردنظر برای جمعیتهای خاص موردنظر محقق در شرایط کاربردی در دنیای واقعی اثربخش خواهد بود یا خیر. در فاز پنجم مطالعه کل افراد جامعه را در نظر می‌گیرند و این موضوع که درمان به چه صورت کل افراد جامعه را تحت تاثیر قرار خواهد داد را مورد بررسی قرار می‌دهند. انجام تمامی این مراحل زمانبر و گران است. برای مثال، برنامه تحقیقاتی فلکمن (Folkman, 1996) بر روی عوامل بازدارنده رشد رگهای جدید در غدد سرطانی برای دهه های متمادی به طول انجامید. این آزمایش با تحقیقات اولیه آزمایشگاهی در مورد نحوه تشکیل رگهای خونی غذارسان به غدده های سرطانی، و اینکه چطور می توان آنها را تحت تاثیر قرار داد شروع شد. نتایج امیدبخش در مدل های حیوانی در اواسط دهه ۱۹۹۰ بدست آمد و زمان بیشتری صرف این شد که دیگر محققین نیز این آزمایش را تکرار کرده و نتایج مشابهی را در دیگر مکانها نیز بدست بیاورند.

اولین آزمایشات دارویی بر روی انسان از سال ۱۹۹۹ شروع شد و دوزهای مختلف دارویی را روی گروه کوچکی از بیماران سرطانی آزمایش کردند تا بتوانند میزان اثربخشی و سمی بودن دوزهای مختلف را بررسی نمایند. نتایج اولیه آن آزمایش ها

[470] Phased models of increasingly generalizable studies

که در اوایل سال ۲۰۰۰ بدست آمد، نشان دهنده اثراتی مثبت (هرچند کوچک) بود. متعاقباً آزمایشهای گسترده انجام شد و احتمالاً تا دهه‌ها ادامه خواهد داشت. این آزمایشها می‌تواند کشف کند که آیا کارایی دارو در کابرد تنهای آن بهتر است و یا در ترکیب آن با دیگر درمانها؟ آیا اثر این درمان روی عوارض مختلف سرطان (مانند زمان مانده تا متاستاز، اثرات خونی، و مدت زمان زنده ماندن بیمار) یکسان است؟ آیا این درمان برای انواع مختلف سرطان به یک اندازه اثربخش است؟
زمینه‌هایی مانند پزشکی سنتهای مستحکم و سابقه‌داری برای حمایت از چنین تحقیقات زمانبر و پرهزینه دارد. اما در دیگر زمینه های تحقیقاتی چنین سرمایه گذاری‌ها و تمایلاتی برای انجام تحقیقات بلندمدت و پرهزینه وجود ندارد. با این وجود، رویکرد فازبندی برای کشف استنباطهای علّی تعمیم یافته، همچنان به عنوان رویکرد برتر مطرح است. این رویکرد بر انتخاب هدفمند نمونه هایی از افراد، موقعیتهای آزمایشی، زمانهای آزمایشی و مداخله ها تکیه داشته و تمامی پنج اصل توصیح علّی را می توان در آن مشاهده کرد، که عبارتند از: (۱) شباهت سطح[۴۷۱]، در بکارگیری سلولهای انسانی سرطانی برای تولید تومورهای با قابلیت مقایسه در حیوانات؛ (۲) خارج کردن یا بی اثر کردن متغیرهای غیرمرتبط از طریق متنوع ساختن بیماران از نظر نوع سرطان، جنسیت و سن بیمار، در اولین دورهای آزمایشهای اثربخشی–سمی بودن؛ (۳) برقرار کردن تمایز از طریق تعیین اینکه درمان در مورد کدام نوع سرطان بیشترین و کمترین اثربخشی را داشته است؛ (۴) دورنیابی و برونیابی در طیفی از دوزهای مختلف دارویی در دورهای اولیه آزمایش برای بررسی اینکه چطور میزان سمّی بودن و پاسخهای کلینیکی برای دوزهای مختلف ممکن است متفاوت باشد؛ (۵) ارائه توضیح علّی برای طراحی مدلهایی که نحوه فعالیت و اثر یک دارو را در از میان برداشتن سلولهای سرطانی توضیح می دهد.

برنامه های هدایت شده آزمایش[۴۷۲]

در این برنامه‌ها با استفاده از آزمایشات متعدد به بررسی سیستماتیک و اکتشافی متغیرهایی (اعم از تعدیلگر، میانجی و دیگر سازه‌ها) که می‌توانند ایجادکننده اثری باشند، پرداخته و از این طریق به طور تدریجی استنباط علّی را تخلیص و دقیقتر می‌نماییم. یک محقق ممکن است این برنامه‌ها را به تنهایی و در طول یک بازه زمانی طولانی در آزمایشگاه انجام دهد، و یا اینکه چندین محقق در آزمایشگاههای متعدد به طور همزمان آنها را انجام دهند. مجموعه‌ای از آزمایشهای انجام شده بر روی مفهوم ابهام مسئولیت نمونه‌ای از اینگونه برنامه‌های هدایت‌شده آزمایشی است. انگیزه‌ی اصلی طراحی این مجموعه از آزمایشها به واسطه حادثه‌ای که در سال ۱۹۶۴ برای زنی به نام کتی گنویس رخ داد بوجود آمد. در این حادثه این زن در انظار عموم برای مدت نیم ساعت مورد ضرب و شتم قرار گرفته و کشته شد. ۴۰ نفر از همسایگان وی در تمام طول این مدت نظاره‌گر این ماجرا بودند، اما هیچکدام به او کمک نکردند. چرا هیچکس با پلیس تماس نگرفت؟ لاتانه و همکارانش (Latane’ & Darley, 1970; Latane’ & Nida, 1981) با انجام بیش از ۳۰ آزمایش نشان دادند که یک توضیح می‌تواند ابهام در مسئولیت باشد. به این معنی که هنگامی که تعداد زیادی از افراد دیگر نیز در محل وجود دارند، افراد فرض می‌کنند که حتماً کس دیگری اقدام خواهد کرد و این فرضیه با افزایش تعداد افراد قوت می‌گیرد. تمامی این ۳۰ آزمایش عناصر

[471] Surface similarity
[472] Directed programs of experiment

نوعی[۴۷۳] مرتبط با مفهوم موردنظر تحقیق را دارا بودند (شباهت سطح). برای مثال، در همه این آزمایشها تماشاچیانی حضور داشتند که قادر بودند به جلوگیری از حادثه ای که در حال رخ دادن بود کمک کنند، موقعیت مورد مشاهده در تمام آزمایشها واجد درجه‌ای از آسیب مشهود و عاجل بود (مثل یک مشکل اورژانسی در سلامتی و یا یک دزدی)، و نتایج همواره عبارت بود از نوعی رفتار کمک کردن متناسب با مشکل. آزمایشها در موقعیتهای متنوعی انجام شدند –زنی که می‌افتد و مچ پایش پیچ می‌خورد، دودی که از پنجره بیرون می‌آید، و یا کسی که دچار حمله صرع می‌شود (اصل بی اثر کردن متغیرهای غیرمرتبط). محققین حتی تعداد تماشاچیان را نیز تغییر دادند، و متوجه شدند که یک تماشاچی که به تنهایی شاهد ماجراست به احتمال زیاد به کمک آسیب دیده می‌شتابد، اما وقتی تعداد تماشاچیان بیش از سه نفر می شود، احتمال کمک کردن به طور قابل توجهی کاهش می‌یابد. این احتمال با افزایش تعداد جمعیت به تدریج (به سمت صفر) کاهش می‌یابد (برونیابی و درون یابی). اگرچه، در صورتی که یک تماشاگر بواسطه نقش اجتماعی خود به طور ویژه مهارت و شایستگی کمک در آن شرایط خاص را داشته باشد (مثلاً یک دانشجوی پزشکی یا یک پرستار اورژانس)، در این حالت تعداد تماشاگران بی‌اثر می شود و چنین فردی همواره (حتی با حضور دیگر تماشاگران) برای کمک کردن پیشقدم می‌شود.

در این مثال، آزمایشها نسبتاً کوتاه (از نظر زمانی)، ارزان و از نظر مراحل انجام آسان بودند. در نتیجه می‌شد آنها را به دفعات تکرار کرد و مطمئن شد که مکانیسم مورد مطالعه قابل تعمیم به موقعیتها، افراد، مداخلهها، نتایج و زمانهای متنوع است. اینگونه برنامه‌ها برای آزمایشهای اجتماعی بزرگتر توجیه‌پذیر نیستند، زیرا آزمایشهای اجتماعی غالباً زمانبر، پرهزینه و از نظر عملیاتی پیچیده هستند. با این وجود، نمونه‌هایی از اینگونه برنامه‌های آزمایشی وجود دارد. الد و همکارانش (Kizman et al., 1997; Korfmacher, O'Brien, Hiatt, & Olds, 1995; Olds et al., 1997; Olds, Henderson, Kitzman, & Cole, 1995) طی چندین مطالعه اثرات سرکشی پرستاران به منزل زنان باردار و نوزادان را مورد بررسی قرار داند. در این برنامه آزمایشی محققین تعمیم‌پذیری اثرات را در میان (۱) زنان آمریکایی–آفریقایی کم‌درآمد شهری، در مقایسه با زنان کم‌درآمد اروپایی–آمریکایی روستایی، (۲) پرستاران در مقایسه با بهیارهای حرفه‌ای، (۳) نتایج سریع در مقابل نتایج مطالعات پیگیری بعد از ۱۵ سال، (۴) متغیرهای نتیجه‌ای متنوع مانند سوءاستفاده از کودکان، سوءتغذیه کودکان، صدمات و جراحات کودکی، فرزندآوری مکرر و رفتارهای مجرمانه و ضداجتماعی کودکان، بررسی کردند. در این مثال بررسی افراد، مداخلهها، نتایج و زمانهای متنوع و تاکید بر یافتن متغیرهای تعدیلگر (هم بی‌ارتباط و هم متمایزکننده)، برای روشن کردن حدود اثر قابل مشاهده است.

لیختن اشتاین (Lichtenstein, Glasgow, & Abrams, 1986) نیز پنج آزمایش ترک سیگار برای دستکاری یک متغیر واسطه‌ای ترتیب دادند. تحقیقات و مشاهدات اولیه محققین را به این فرضیه رساند که میزان حمایت اجتماعی که بیمار دریافت می‌کند، اثر مداخلات ترک سیگار را میانجی‌گری می‌کند. بنابراین در این پنج مطالعه، محققین برنامه ترک سیگار شناختی–رفتاردرمانی را در حالتهای با و بدون عنصر حمایت اجتماعی مقایسه نمودند. اگرچه چهار از پنج مطالعه نشان‌دهنده وجود همبستگی میان حمایت اجتماعی و نتایج بود، خود دستکاریها تفاوتی میان شرایط وجود و عدم وجود حمایت نشان نمی‌داد. محققین تعدادی تبیین یا توضیح محتمل برای این شکست ارایه نمودند، از قبیل اینکه در دستکاریهای

[473] prototypical

صورت‌گرفته سطح حمایت اجتماعی به اندازه کافی بالا نبوده است تا بتواند نتایج را تحت تاثیر قرار دهد؛ یا اینکه در صورتی که تحلیلها به تفکیک گروههای جنسیتی و یا به تفکیک سطوح نیاز به حمایت اجتماعی انجام می شد، ممکن بود اثرات معنادار شوند.

مرور روایتی[۴۷۴] تحقیقات موجود

با در نظر گرفتن زمان و هزینه لازم برای انجام برنامه‌های تحقیقاتی با قابلیت تعمیم‌یابی، یک استرتژی جایگزین، می‌تواند مرور ادبیات موجود برای یافتن نشانه‌هایی از استنباطهای علّی تعمیم‌یافته باشد. مزایا و معایب چنین مرورهایی نقطه مقابل برنامه‌های تحقیقاتیست. به این معنی که این تحقیقات زمان و هزینه بسیار کمتری می‌گیرند، البته محقق را به متغیرها و سوالاتی محدود می‌کنند که (تا به حال) در ادبیات موجود مورد بررسی قرار گرفته است. در این بخش به توضیح مرورهای روایتی، و در بخش بعدی به مرورهای کمّی خواهیم پرداخت.

مرور روایتی آزمایشها

این مرورها ادبیات موجود را با استفاده از تبیینهای روایتی، و بدون هیچ‌گونه تلاشی برای سنتز کمّی نتایج مطالعات (که به آن متاآنالیز گفته می‌شود)، توضیح می‌دهند. برای مثال، گورمن و همکارش (Gurman & Kniskern, 1978) دویست مطالعه صورت گرفته در مورد اثرات رواندرمانی خانواده و زوجها را مرور نموده، و با استفاده از مقادیر متنابهی متن و جدول، مشخصات این مطالعات و جهت‌های کلی نتایج آنها را توضیح می‌دهند. مطالعات مورد بررسی در مرور موردنظر توسط محققین مختلف، و با استفاده از روش‌شناسیها و مقیاسهای ارزیابی نتایج مختلف، بر روی انواع مختلفی از مراجعان، با بکارگیری ویرایشهای مختلفی از درمانهای خانواده و زوج، و با طول درمانهای مختلف انجام شده بود. این دو محقق بر اساس ارزیابی اینکه آیا افرادی که درمانهای خانواده و زوج دریافت کرده‌اند به طور معناداری نسبت به گروهی که درمانهای کنترل را دریافت کرده‌اند، عملکرد بهتری داشته‌اند، چنین نتیجه‌گیری کردند که این درمانها به طور کلی اثربخش بوده‌اند. برای مثال، آنها دریافتند که در ۱۸ مطالعه افراد دریافت‌کننده درمان خانواده و زوج عملکرد بهتری از گروه کنترل داشتند، در ۱۱ مطالعه تفاوتی در عملکرد دو گروه مشاهده نشد، و در ۲ مطالعه عملکرد گروه درمان ضعیفتر از گروه کنترل بود. به همین ترتیب، گروههای درمان دریافت‌کننده‌ی زوج–درمانیهای رفتاری، در ۷ مطالعه از ۱۱ مطالعه عملکرد بهتری از گروه کنترل نشان می‌دادند، و دریافت‌کنندگان خانواده‌درمانی رفتاری در تمامی ۵ مطالعه مرتبط، نتایج بهتری را نسبت به گروه کنترل کسب کرده بودند. این شکل از شمارش موفقیتها، تساویها، و شکستها (در ادبیات موجود)، اغلب رویکرد نمرات باکس[۴۷۵] یا شمارش رای در مرور مطالعات اثربخشی نامیده می‌شوند (Hedges & Olkin, 1985). محقق مرورکننده برای نتیجه‌گیری در مورد اثربخشی مداخلات بر سطوح معناداری آماری تکیه می‌کند؛ به اینصورت که اگر گروه درمان به طور معناداری بهتر از گروه کنترل عمل کرده باشد، یک رأی مثبت (یا موافق)، اگر ضعیفتر عمل کرده باشد یک رای منفی (یا

[474] Narrative review of existing research
[475] Box Score

مخالف)، و اگر دو گروه تفاوت معناداری نداشته باشند یک رای تساوی در نظر می‌گیرد. بر اساس جمعبندی این رأیها پیشنهاد خواهد کرد که آیا درمان موردنظر در طول مطالعات مختلف اثربخش بوده یا خیر.

چنین مروری همچنین امکان بررسی متغیرهای مداخله‌گر بالقوه اثرگذار بر تعمیم‌پذیری اثر مداخلات را فراهم می‌آورد. برای مثال، گورمن و همکارش چنین گزارش می‌کنند که درمانهای کوتاه‌مدت از اثربخشی برابر با درمانهای بلندمدت برخوردارند و اینکه عملکرد مراجعین در حضور درمانگران با تجربه و بی‌تجربه به یک میزان است– در نتیجه اثر در سطوح مختلف دوز درمان و تجربه درمانگر تعمیم داده شد. عدم توفیق در تعمیم نیز می‌تواند بسیار جالب باشد. برای مثال، زوج–درمانی هنگامی بهتر جواب خواهد داد که زن و شوهر هر دو در درمان مشارکت نمایند (در مقایسه با زمانی که تنها یکی از آنها درمان را دریافت می‌کند). همچنین در مورد بیماران مضمن و حاد بیمارستانی، تنها مداخلات مبتنی بر نظریه سیستمها اثربخش گزارش شده بود و نه هیچ درمان دیگری.

مرورهای روایتی در که در آنها تحقیقات آزمایشی و غیرآزمایشی ترکیب می‌شوند

مرورهای روایتی می‌تواند نه فقط آزمایشهای میدانی را، بلکه شواهد بدست‌آمده از مطالعات پیمایشی، مطالعات صورت گرفته بر حیوانات، کارهای آزمایشگاهی پایه را نیز لحاظ کنند (در بر بگیرند) تا شواهد موجود در این مطالعات را با الگویی که بیانگر وجود اثر و تعمیم‌پذیری آن است تطبیق دهند. برای مثال، دویر و همکارش (Dwyer & Flesch-Janys, 1995) الگویی از شواهد را تجمیع و خلاصه‌سازی نمودند که می‌توانست رابطه علّی میان استفاده از گاز نارنجی در جنگ ویتنام و بروز متعاقب سرطان در افرادی که در معرض این گاز بوده اند را پشتیبانی نماید. برخی مطالعات نشان داد که سطح سمیّت بافتها و خون در افرادی که در محیطهای آلوده به این گاز حضور داشتند شش برابر بیشتر از افرادی بود که در محیطهای غیرآلوده بودند. در مرحله بعدی، این دو محقق آزمایشهای صورت‌گرفته بر روی حیوانات و مطالعات اپیدمیولوژی انسانی را بکار گرفتند تا نشان دهند که این سم با افزایش سرطان همراه است– حتی در حیواناتی که بیشترین مقاومت را در مقابل سموم قوی دارند و در افرادی در آمریکا و آلمان که سطح مواجهه ای بالاتر از حد عادی با این گاز داشتند (در آلمان به این دلیل که این گاز در کارخانجات این کشور تولید می‌شد). نتایج در میان انواع و مکانهای مختلف سرطان صادق بود (مصداق داشت). این دو محقق همینطور نوعی تجانس در شواهد موجود مرتبط با مکانیسم علّی‌ای که از طریق آن گاز پرتغالی می‌تواند ایجادکننده سرطان در انسان و حیوان باشد را نشان دادند. با ترکیب این منابع چندگانه شواهد، این دو محقق دلایلی قانع‌کننده مبنی بر موجّه بودن[۴۷۶] (توجیه پذیربودن) رابطه میان گاز پرتغالی و سرطان در ویتنام، توضیحاتی برای این رابطه و تعمیم‌پذیری آن ارایه نمودند.

اینگونه مرورها خصوصاً زمانی بیشتر به کار می‌آیند که دستکاری مستقیم آزمایشی بر روی انسان غیراخلاقی تلقی شود (مانند مثال گاز پرتغالی) یا غیرممکن باشد (مانند جنسیت یا رویدادهای گذشته). در چنین مواردی، دستکاری آزمایشی حیوانات، تحقیقات پایه‌ای آزمایشگاهی و دانش نظری موجود را می‌توان با شبه–آزمایشها و داده‌های مشاهده‌ای بدست‌آمده از تنوعهایی که به طور طبیعی در درمان رخ می‌دهند ترکیب نمود تا یک اثر و میزان تعمیم‌پذیری آن را مشخص نمود.

[476] Plausibility

اگرچه، با در نظر گرفتن نقدهایی که بر نقایص روش‌شناختی اینگونه مطالعات وارد می‌آید، نتایج این مرورها می‌تواند بحث‌برانگیز باشد. رابطه علّی میان مصرف سیگارو سرطان نمونه‌ای از این مسأله است. به دلایل اخلاقی، غالب تحقیقات صورت گرفته در این خصوص از نوع همبستگی و شبه–آزمایشی هستند. با وجود اینکه این رابطه امروزه کاملاً پذیرفته‌شده به نظر می‌رسد، دانشمندانی مانند سر رونالد فیشر و هانس ایزنک (Grossarth-Maticek & Eysenck, 1989; Pearl, 2000) آن را به چالش می‌کشند. به همین ترتیب، امروزه رابطه میان تدخین ثانویه (دست دوم) و سرطان مورد مناقشه است؛ بحثها عمدتاً از ابهامات روش‌شناختی نشات می‌گیرند.

ایرادات مرورهای روایتی

مرورهای روایتی نقاط قوت مهمی دارند که معادل آن در روشهای کمّی مرور که در ادامه به آنها خواهیم پرداخت یافت نمی‌شود (به دلیل تولید فرضیات، تبیین حجیم ادبیات، و تولید نظریه بر مبنای طبقه‌بندیهای کیفی و روابط میان متغیرها). اما از جهت امکان ارزیابی استنباط علّی مورد تعمیم، این مرورها دچار ضعفهایی هستند.

نخست، با افزایش یافتن تعداد مطالعات، سازماندهی تمامی روابط میان اثرات و متغیرهای مداخله‌گری که می‌توانند مهم باشند، به طور تصاعدی دشوار می‌شود. اینکار به سختی حالتی است که حین انجام مطالعات اولیه تلاش می‌کنیم تا تمامی داده‌های بدست‌آمده از تمامی پاسخ‌دهندگان را با استفاده از توضیحات روایتی تخلیص نماییم. تا حدی هم به این دلیل است که از اعداد برای نشان دادن مشاهدات استفاده می‌کنیم. اعداد کمک می‌کنند تا بتوان داده‌ها را به درستی درک نموده، آنها را به صورت کاراتری سازماندهی کرده و رابطه میان داده‌ها را به صورت آماده‌تری تحلیل کرد. بدون تردید، به طور اجتناب ناپذیری چیزهایی در جریان حرکت از تبیین روایتی به سمت اعداد از دست می‌روند. این باعث می شود تا بسیاری از محققین پیشگام هر دو روش را در مطالعات ابتدایی مورد استفاده قرار دهند (Reichardt & Rallis, 1994).

دوم، مرورهای روایتی به طور سنتی بر خلاصه‌های نمرات باکس حاصل از آزمونهای معناداری نتایج مطالعات تکیه دارند. اما این اطلاعات بسیار محدود است. یک اثر معنادار به این معناست که کمتر احتمال دارد تفاوتهای مشاهده شده میان گروهها بواسطه شانس در جمعیتی که در آن، اثر مداخله وجود نداشته باشد، رخ داده باشد. این معناداری اطلاعات ناچیزی در مورد اندازه و بزرگی تفاوت میان گروه‌ها بدست می‌دهد. تفاوت بسیار کوچکی میان گروهها، در مطالعه‌ای که دارای نمونه‌ای بسیار بزرگ است، می‌تواند معنادار شود؛ حتی اگر این اندازه تفاوت هیچ اهمیت کاربردی خاصی نداشته باشد. یا در مقابل، اثرات بسیار بزرگ در مطالعات با اندازه نمونه کوچک غیرمعنادار شوند، حتی اگر این اثرات از نظر کاربردی مهم باشند (Kazdin & Bass, 1989; Mosteller et al., 1980). بنابراین به هنگام انجام مرور در مورد اثر یک مداخله، اطلاعات مرتبط با اندازه اثر بسیار ارزشمند تلقی می‌شوند.

سوم، مرورهای روایتی در توضیح نتایج مطالعه چندان دقیق نیستند. فرض کنید یک مرورکننده از رویکرد نمرات باکس استفاده می‌کند، و پیشنهاد می‌کند ۵ مورد از میان ۱۰ مطالعه انجام شده نتایج غیرمعنادار تولید کرده‌اند. این خلاصه می‌تواند حاصل یکی از دو الگو باشد. در یکی، پنج مطالعه‌ی با نتایج معنادار ممکن است خطای نوع اول کمتر از ۰۰۱/ (p<.001) گزارش کرده باشند، و پنج مطالعه غیرمعنادار p=.10؛ و در همه موارد نمونه‌های کوچک مورداستفاده قرار گرفته است. در

الگوی دوم، فرض کنید در پنج مورد از نتایج معنادار به لحاظ آماری p=.04 و در نتایج غیرمعنادار، خطای نوع اول نرخی مابین p=.5 و 99. دارد؛ و تمام آزمونها بر اساس نمونه‌های بزرگ بدست آمده‌اند. با رویکرد نمرات باکس در هر دو این موارد نمرات مشابهی به دست می‌آوریم: ۵ رای موفقیت و ۵ تساوی. اما تجمیعی دقیقتر از نتایج می‌توانند مشخص‌کننده اثری بسیار بزرگ در الگوی اول، و اثری کوچک در الگوی دوم باشد.

چهارم، زمانی که مرورکننده در صدد برمی‌آید تا روابط میان نتایج و متغیرهای مداخله‌گر احتمالی را مورد بررسی قرار دهد، مسایل بیش از پیش پیچیده می‌شود. در این حالت مرورکننده باید با ناراستیها و بی‌دقتی‌های موجود در متغیر مداخله‌گر، پایش تعداد بیشتری رابطه میان متغیرها، و این سوال که آیا اندازه تفاوتها در سطوح مختلف متغیر مداخله‌گر و دیگر متغیرها تغییر می‌کند، نیز مواجه شود. با وجود چنین مشکلاتی، مرور روایتی ادبیات مداخلات می‌تواند بسیار دشوار باشد.

بنابراین جای تعجب نیست که مرورهای روایتی ادبیات هر چه بیشتر جای خود را به مرورهایی که ترکیبی از روشهای کمّی و کیفی هستند می‌دهد (C. Mann, 1994). اینگونه مرورها به همان شیوه و به همان دلایلی از اعداد برای توضیح مطالعات اولیه استفاده می‌کنند، که مطالعات اولیه[۴۷۷] از اعداد برای توضیح شرکت‌کنندگان خود بهره می‌گیرند. اما بکارگیری اعداد به معنی رد تحلیلهای کیفی یا روایتی نیست. دقیقاً همانطور که مطالعات اولیه از بکارگیری روشهای کمّی و کیفی منتفع می‌شوند (Reichardt & Rallis, 1994)، مطالعات مرور بر ادبیات نیز از هر دو رویکرد سود خواهند برد.

مرور کمّی تحقیقات موجود

بکارگیری تکنیکهای کمی برای خلاصه کردن نتایج مطالعات علمی سابقه‌ای طولانی دارد. در اوایل قرن ۱۸، ریاضی‌دانی انگلیسی به نام راجر کوته میانگین وزنی مقیاسهای بدست آمده توسط ستاره‌شناسان متفاوت را محاسبه کرد. کارل پیرسن (Sir Karl Pearson, 1904) روشهای کمّی را برای یافتن نتایج شش مطالعه در مورد اثرات یک راه درمانی جدید برای درمان تیفوئید بکار گرفت. با این حال، این روشها تا پیش از معرفی مفهوم متاآنالیز توسط گلس (Glass, 1976) با اقبال اندکی مواجه بودند. واژه متاآنالیز به عنوان مفهومی برای تبیین تکنیکهای کمّی محاسبه اندازه اثر در میان مطالعات متعدد بکار می‌رود. اگرچه پیش از گلس نیز دیگر محققین روشهای کمّی را برای مرور بر ادبیات به کار گرفته بودند، اما گلس اولین کسی بود که روشی جامع را برای تجمیع اندازه اثرهای بدست‌آمده از مطالعات مختلف طراحی کرد.

نوآوری اساسی استفاده از اندازه اثر به عنوان شاخص اندازه‌گیری مشترک در میان مطالعات است. وجود یک شاخص اندازه‌گیری مشترک از آنجا اهمیت دارد که مطالعات به ندرت از مقیاسهای یکسانی استفاده می‌کنند، حتی زمانی که به سوال مشابهی پاسخ می‌دهند و مفاهیم مشابهی را مورد بررسی قرار می‌دهند. بنابراین، مثلاً برای ارزیابی افسردگی ممکن است در مطالعه‌ای از مقیاس افسردگی بک استفاده شود، در حالی که مطالعه‌ای دیگر مقیاس MMPI را برای محاسبه سطح افسردگی بکار گیرد. این شاخص‌های اندازه‌گیری مقیاسهای متفاوتی دارند که میانگینها و انحراف معیارهای متفاوتی به دست می‌دهد. بنابراین، محاسبه میانگین آنها پیش از آن که مقیاس‌های آنها یکسان شود، کاری بیهوده و غیرمنطقی خواهد بود. متاآنالیز نتایج این قبیل مطالعات را به یک مقیاس مشترک اندازه اثر تبدیل می‌کند، به گونه‌ای که نتایج مختلف

[477] Primary studies

میانگین‌ها و انحراف معیارهای یکسانی داشته باشند و با سهولت بیشتری بتوان متوسط آنها را در میان مطالعات مختلف محاسبه نمود[۴۷۸]. اندازه اثر در ابتدا برای محاسبه اثرات در حوزه رواندرمانی و اثر اندازه کلاس بر موفقیت تحصیلی دانش آموزان بکار گرفته شد. اما امروزه متاآنالیز در مورد اثربخشی مداخلهها و مداخلات در بسیاری از زمینه‌ها، از جمله پزشکی، اقتصاد کار، و حشره شناسی بکار می‌رود. البته اصول مرور کمّی محدود به بررسی اثر مداخلات نیست. بلکه متاآنالیز برای سوالات تحقیقی که واجد مداخله نبوده‌اند نیز بکار گرفته شده است. مثلاً برای بررسی اینکه آیا دختران و پسران از نظر موفقیت علمی متفاوت بوده‌اند یا خیر؛ و یا برای بررسی روایی آزمونهای استفاده شده در انتخابهای شخصی؛ و یا برای بررسی متغیرهای همبسته اثر انتظارات بین فردی. همچنین می‌توان متاآنالیز را برای سوالات کاملاً توصیفی مانند درصد بیماران روانی قبلی که بعد از مرخص شدن از تیمارستان مرتکب جنایت شدند نیز بکار گرفت. در این مورد آخر، استفاده از عبارت «اندازه اثر» باید به طور گسترده به عنوان شاخصی کمّی از بزرگی و قدرت در نظر گرفته شود.

مبانی متاآنالیز

در این قسمت به مرور قدمهای ضروری برای انجام متاآنالیز و مفاهیم پایه‌ای لازم برای فهم بهتر این روش خواهیم پرداخت. شناخت درست این مفاهیم و روشها به خواننده در درک نحوه بکارگیری متاآنالیز در استنباطهای علّی تعیم‌یافته کمک می‌نماید. اگرچه خوانندگانی که قصد انجام متاآنالیز را دارند، باید دیگر منابعی که روش را با جزئیات بیشتر توضیح می‌دهند را مطالعه کنند. مرور بر ادبیاتها خواه به صورت کیفی و خواه کمّی، مراحلی را طی می‌کنند. این مراحل عبارتند از تعریف مسأله، جمع‌آوری، ارزیابی، و تجزیه و تحلیل داده‌ها، و ارائه گزارش از نتایج. به مانند هر تحقیق دیگری، این مراحل چرخشی و تکرارپذیر[۴۷۹] بوده و تا حد زیادی به یکدیگر وابسته‌اند (و اینطور نیست که قدمهای مجزایی در هر واحد زمان انجام شده و پایان یابد). در بخش بعدی برخی از رویکردهای مختص متاآنالیز و ملزومات همراه با آنها را در مورد مراحل مشترک تحقیقات ارائه می‌کنیم.

شناسایی مسأله تحقیق و انجام مرور بر ادبیات

هر مروری بر ادبیات باید با توجه به مسأله مورد تحقیق صورت پذیرد. مسأله مورد بررسی می‌تواند بسیار وسیع باشد، مانند ارزیابی اثرات کلی رواندرمانی؛ و یا به صورت جزئی تعریف شده باشد، مانند ارزیابی اینکه آیا بکارگیری مداخلات متناقض در رواندرمانی می‌تواند بهبوددهنده نتایج باشد. مسأله می‌تواند به اثر اصلی یک مداخله در مقایسه با گروههای کنترل و یا دیگر مداخلهها بپردازد؛ یا اینکه می‌تواند اثرات متغیرهای مداخله‌گری مانند اینکه آیا اثرات رواندرمانی در مراحل پس از

---

[۴۷۸] یک راه جایگزین ترکیب سطوح احتمال است. مطالعات مداخله–محور عموماً معناداری تفاوت میان گروه آزمون و گروه کنترل را با در نظر گرفتن سطح احتمال بروز خطای نوع اول می سنجد. روشهای مختلفی برای تجمیع سطوح احتمال وجود دارد و نتیجه آنها آزمون معناداری ترکیبی بدست می‌دهد که فرض صفر آن عبارت است از اینکه اثر مفروض در هیچکدام از جمعیتهای بررسی شده در مطالعات مختلف معنادار نیست؛ و فرض مقابل به این صورت بیان می‌شود که حداقل یکی از نمونه‌های بررسی شده از جمعیتی با اثر غیرصفر گرفته شده است (اثر در آن مشاهده شده است). اغلب محققین ترجیح می‌دهند که درباره اندازه اثر اطلاعات داشته باشند، تا اینکه به راحتی به دانستن اینکه حداقل در یکی از مطالعات اثر غیرصفر بدست آمده بسنده کنند. اگرچه این روشهای سطح معناداری ترکیبی هنگامی مفید هستند که اطلاعات کافی برای محاسبه اندازه اثر وجود نداشته باشد اما سطوح معناداری در دسترس باشد. وانگ و بوشن (Wang & Bushman, 1999) روشی را برای ترکیب اندازه اثرها و سطوح احتمال ارائه می‌نمایند.

[۴۷۹] Iterative

آزمون و آزمونهای پیگیری همچنان باقی می‌ماند یا خیر را مورد توجه قرار دهد. مسأله همچنین می‌تواند به یافتن واسطه‌ای برای اثرات مشاهده شده بپردازد. فرایند طراحی و فرمولبندی سوال تحقیق در متاآنالیز تفاوت چندانی با فرایند تدوین سوال تحقیق به طور عمومی ندارد. هدف طراحی یک سوال تحقیق روشن است. سوالی که بر مبنای آن بتوان یک چهارچوب قابل تغییر از معیارها، برای لحاظ کردن یا مستثنی کردن مطالعات در فرایند متاآنالیز (مرور بر ادبیات) بدست آورد.

پس از تدوین سوال تحقیق لازم تا محقق به جستجوی تحقیقات مرتبط بپردازد. بسیاری از تحقیقات انجام شده را می‌توان از پایگاههای داده، بررسی منابع مرورهای انجام شده قبلی، مرور فهرست مجلات اخیر، شناسایی آزمایشهای انجام شده و در حال انجام و همچنین برقراری ارتباط با دیگر محققین فعال در حوزه مربوطه بدست آورد. اما ممکن است دسترسی محدودی به برخی از مطالعات وجود داشته باشد. مثلاً نتایج منتشر نشده، پایان‌نامه‌های کارشناسی ارشد و دکتری، گزارشهای نهایی طرحهای پژوهشی سازمانها، گزارشهای فنی و مطالعاتی که شانس چاپ شدن را از دست داده‌اند و سر از کشوی محقق در آورده‌اند. مقدار این گونه مطالعات مشخص نیست. اما در برخی زمینه‌های تحقیق ممکن است این مقدار قابل توجه بوده و اثرات کوچکتری را نسبت به اثرات گزارش شده در تحقیقات منتشر شده در همان زمینه گزارش نمایند. اگر چنین موقعیتی وجود داشته باشد، ضرورت دارد تا اینگونه تحقیقات (با فرض اینکه از نظر معیارهای تحقیق قابلیت لحاظ شدن در تحقیق را داشته باشند) در مرور بر ادبیات لحاظ شود تا حذف نتایج آنها منجر به سوگیری در تخمین اثرات نشود. روزنتال (Rosenthal 1994) روشهایی را برای یافتن و در نظر گرفتن تحقیقات منتشر نشده ذکر می‌کند، اگرچه محقق هیچگاه نمی‌تواند مطمئن باشد که تمامی آنها را یافته است.

مطالعاتی که در فرایند مرور بر ادبیات لحاظ می‌شوند باید معیارهای اساسی مطالعه را داشته باشند. آنها می‌بایست سوال موردنظر تحقیق را پاسخ داده و به مداخلهها، افراد، مشخصات آزمایش، مقیاسها و زمانهای عنوان شده در مسأله تحقیق مربوط باشند. ممکن است معیارهای روش‌شناختی نیز برای لحاظ کردن مطالعات در فرایند مرور در نظر گرفته شود. برخی از این معیارها ممکن است با یکدیگر در تضاد باشند؛ برای مثال این که آیا لازم است محدودیتی برای اندازه نمونه و توان آماری آزمون در هر مطالعه در نظر بگیریم (Kraemer, Gardner, Brooks,& Yesavzge, 1998; Wortman, 1992)، آیا تنها مطالعات چاپ شده را در نظر بگیریم، و اینکه آیا تنها مطالعات تصادفی را لحاظ کنیم (هنگامی که مرور بر ادبیات در مورد یک مساله علّی است) (Boissel et al., 1989). اینگونه تصمیمات تا حد زیادی به زمینه انجام متاآنالیز بستگی دارد. برای مثال، به دلایل اخلاقی مطالعات در مورد مراقبتهای نوزادی ندرتاً به صورت تصادفی انجام می‌شوند و بنابراین متاآنالیزها به شبه‌آزمایشها محدود می‌شوند (Ozminkowski, Wortman, & Roloff, 1989). در دیگر مواقع، ممکن است تعداد آزمایشات تصادفی چنان زیاد باشد که تمام منابع موجود مصروف کدگذاری آنها شود (Shadish et al., 1993).

هنگامی که در اتخاذ این تصمیمات دچار تردید می‌شویم، ترجیح بر آن است که تعداد بیشتری از مطالعات را لحاظ نموده و سپس معیارهای روش‌شناختی را کدگذاری کنیم، تا در مراحل بعدی بتوان اثر احتمالی آن را در نتایج مورد بررسی قرار داد.

چه تعداد مطالعه با یک سوال مشترک باید وجود داشته باشد تا بتوان یک متاآنالیز را به انجام رساند؟ پاسخ این سوال به مساله «توان آماری» ارتباط پیدا می‌کند. در متاآنالیز محاسبه توان به روشهای استاندارد مرسوم که در مطالعات تکی انجام می‌شود، صورت نمی‌پذیرد. بلکه روشهای خاص توان متاآنالیز باید واریانس بین مطالعات، توان برخی آزمونها که برای

متاآنالیز ناآشنا و غیرمعمول باشد (مانند آزمون یکدستی یا تجانس[۴۸۰])، و توان متفاوت مدلهای اثرات ثابت در مقابل تصادفی، را در نظر بگیرند. هرج و اوکلین (Hedges & Oklin, chapter7) معادلات پیش فرض برای این موارد را معرفی می‌نمایند.

کدگذاری مطالعات

متاآنالیزها یک چهارچوب مشترک کدگذاری را برای کمّی کردن نتایج و مشخصات مطالعات تحت بررسی بکار می‌گیرند. مهم آن است که کدها باید نشان‌دهنده فرضیات محقق باشند. برای مثال، در یک مطالعه روی مداخلات انجام شده درباره چاقی کودکان و نوجوانان، ممکن است محقق نسبت به این موضوع کنجکاو باشد که آیا مداخلات شامل مولفه‌های رفتار آموزشی، ورزشی، و رژیم غذایی بوده‌اند یا نه، یا اینکه آیا نسبت به سن کودکان کنجکاو بوده، و یا اینکه مایل است این مسأله را بررسی کند که آیا والدین در جریان مداخلات مشارکت داشته‌اند یا خیر. برای هرکدام از این اهداف، کدهایی باید طراحی شوند. کدها اغلب دربرگیرنده مشخصات گزارش مطالعه (برای مثال تاریخ انتشار، شکل انتشار)، مشخصات مداخله یا مداخله (مثلاً استفاده از دستورالعمل برای افزایش پایبندی به مداخلات، سطح پایبندی بدست آمده)، و مشخصات روش‌شناختی (مانند اندازه نمونه، روش تخصیص به شرایط آزمون و کنترل، ریزش و انواع مقیاسها) هستند.

معمولاً انجام این کدگذاریها سخت‌تر از آن است که به نظر می‌رسد. یکی از دلایل آن، وجود ابهامات و اهمال‌های محتمل در گزارش‌دهی‌های اولیه است (Orwin & Cordray, 1985; Pigott, 1994). برخی اوقات می‌توان این مشکل را از طریق برقراری تماس با نویسندگان مقالات مرتفع کرد. اما پیدا کردن نویسندگان مطالعات قدیمی کار آسانی نیست؛ ضمن آنکه ممکن است مدارک مرتبط با مطالعه را حفظ نکرده باشند و حافظه آنها نیز یارای چندانی در این خصوص ننماید. برخی کدها وابسته به دیگر کدها هستند. برای مثال، ممکن است دو کد به طور همزمان یک اندازه اثر را برای یک متغیر وابسته محاسبه نمایند، اما اگر یکی متغیر را به یک صورت (مثلا به صورت واکنشی) و دیگر به صورتی متفاوت (مثلاً غیرواکنشی) محاسبه کرده باشند، پایایی کلی اندازه اثر برای مقایسه واکنشی–غیرواکنشی کاهش می‌یابد (Wortman, 1992; Yeaton & Wortman, 1993). علاوه بر این برخی کدگذاری ها نیازمند دستورالعمل‌های مفصل (همراه با جزئیات) و قضاوتهای فصل‌بندی شده است؛ مانند زمانی که میزان واکنش یک مقیاس درجه‌بندی می‌شود، و یا زمانی که جهت‌گیری نظری یک روش رواندرمانی طبقه‌بندی می‌شود. از این رو، چهارچوب شماتیک اولیه باید قوانین کدگذاری روشنی را دنبال کند، باید از نظر پایایی اینتررییتر[۴۸۱] آزمون شود، و باید تا زمانی که پایایی موردنظر حاصل شود، اصلاح و بازبینی شود. آزمونهای پایایی دوره‌ای و بازآموزی‌های مستمر گزینه مناسبی برای تضمین کردن عدم انحراف کدگذاری در طول زمان هستند (Orwin, 1994).

در نهایت، پروتکل کدگذاری باید با در نظر گرفتن نحوه ورود داده برای تجزیه و تحلیلهای آماری بعدی، و اینکه آیا کدگذاریها مستقیما بر روی فایلهای کامپیوتری انجام می‌شود یا نه، طراحی شود (Woodworth, 1994). مشکلات همراه با وارد کردن داده در مطالعات متاآنالیز غالباً پیچیده‌تر از مطالعات اولیه است. زیرا عموماً سطوح متعددی از لانه‌گزینی در

[480] Homogeneity
[481] intermeter

متاآنالیز وجود دارد (برای مثال، اندازه اثرها در درون مقیاسها، در زمانهای مختلف، در درون مقایسه‌های مابین مداخلات در مطالعات). این مسأله عموماً منجر به شکل گرفتن فایلهایی با ساختار پیچیده می‌شود و باید در هنگام تدوین پروتکل کدگذاری و ورود داده‌ها در نظر گرفته شود. لیپسی و ویلسن (Lipsey & Wilson, 2000) توصیه‌های مفیدی در مورد این ساختارها ارائه می‌نمایند.

محاسبه اندازه اثر

به ندرت اتفاق می‌افتد که مقیاس یکسانی برای اندازه‌گیری نتایج در همه یا بخشی اعظمی از مطالعات تحت بررسی در متاآنالیز بکار گرفته شده باشد. متاآنالیز اثرات مداخلات آموزشی–روانی قبل از جراحی برای بهبود نتایج پس از جراحی می‌تواند شامل مطالعاتی باشد که مقیاسهای مختلف متعددی را برای ارزیابی درد پس از جراحی بکار گرفته باشند (مانند تعداد قرصهای ضددرد درخواست شده از سوی بیمار، تعداد شکایات به پرستار، و گزارش شخصی فرد بر اساس یک مقیاس درجه‌بندی). ممکن است این چنین مطالعات و مفاهیم سازه‌های متفاوتی را اندازه‌گیری کنند (مثلاً نه تنها درد پس از جراحی، بلکه طول مدت ماندن در بیمارستان، رضایتمندی مشتری و اضطراب ناشی از جراحی را اندازه‌گیری کنند). در نتیجه، قبل از اینکه بتوان نتایج را به طور معناداری با یکدیگر مقایسه نمود، نتایج مختلف مطالعات باید به یک مقیاس مشترک تبدیل شود. این مهم طی محاسبه اندازه اثر محقق می‌شود. روشهای مختلفی برای اندازه‌گیری اندازه اثر امکان‌پذیر است (Fleiss, 1994; Rosenthal, 1994). در اینجا به دو مورد از مناسب‌ترین آنها برای متاآنالیز آزمایشها اشاره می‌کنیم.

- آماره تفاوت میانگین استاندارد (d)
- نرخ غیرعادی‌ها (o)

آماره تفاوت میانگین استاندارد شده به صورت زیر تعریف می شود:

$$d_i = \frac{\bar{x}_i^t - \bar{x}_i^c}{S_i}$$

بطوریکه $\bar{x}_i^t$ میانگین گروه آزمون در مطالعه i ام است. $\bar{x}_i^c$ میانگین گروه مقایسه در مطالعه i ام است. و S_i انحراف معیار تجمعی دو گروه است. S_i به راحتی از طریق دانستن اندازه نمونه و انحراف معیارهای دو گروه در حال مقایسه بدست می‌آید. در متاآنالیزهایی که دربرگیرنده مطالعات با اندازه نمونه کوچک هستند، برای برطرف کردن سوگیری نمونه کوچک باید به طور مستمر تصحیح صورت بگیرد (Hedges & Olkin, 1985, p.81). گروه مقایسه شونده می‌تواند یک گروه کنترل و یا یک گروه دریافت‌کننده آزمون جایگزین باشد.[۴۸۲] برای مثال فرض کنید که یک روش خانواده‌درمانی برای زوجهای دچار مشکل، نمرات پس–آزمون با میانگین ۱۰۳/۹۷ در مقیاس رضایتمندی زناشویی تولید کرده، در حالی که گروه کنترل

۴۸۲– ضریب همبستگی یکی دیگر از مقیاسهای اندازه‌گیری است. این مقیاس زمانی بیشترین کارایی را خواهد داشت که متاآنالیز مطالعاتی را جمعبندی و خلاصه کند که روابط همبستگی یکسانی را میان متغیرها آزمون کرده‌اند. بیشتر محققین پیشنهاد می‌کنند قبل از تبدیل z، پایدارکننده واریانس فیشر (۱۹۵۲) استفاده شود. در تحقیقات دارای مداخله، همبستگی میان نمرات دوتایی شرایط مداخله و کنترل (مداخله= ۱ و کنترل= ۰) با یک مقیاس پیوسته در نتایج (مثل سطح افسردگی، تعداد روزهای ماندن در بیمارستان) می‌تواند شکل‌دهنده مقیاسی برای اندازه‌گیری اندازه اثر باشد. اگرچه به ندرت این همبستگی در مطالعات مداخله گزارش می‌شود.

میانگینی برابر با ۹۷/۱۳ در همین مقیاس داشته‌اند. اگر انحراف‌معیار تجمعی برابر با ۲۲/۰۶ باشد، اندازه اثر (d) برابر با ۰/۳۱ خواهد بود. به این معنی که در پس–آزمون گروه مداخله به میزان یک سوم انحراف معیار بهتر از گروه کنترل عمل کرده است. هنگامی که مقادیر میانگین، انحراف معیار و اندازه نمونه در اختیار نباشد، اغلب امکان دارد که دیگر آماره‌های گزارش شده در مطالعه را برای محاسبه (d) یا تخمین تقریبی از d مورد استفاده قرار دهیم. اگرچه تحقیق چندانی درباره مشخصات این تقریبها صورت نگرفته است. بسیاری از آنها کاملاً خوب عمل می‌کنند اما معدودی از آنها ممکن است تقریبهای ضعیفی از روش اصلی ارائه کنند (Shadish et al., 1999).

آماره تفاضل میانگین استانداردشده فرض را بر این می‌گذارد که نتایج مطالعه به طور پیوسته اندازه‌گیری شده است. اما برخی اوقات متغیر نتیجه روی مقیاس دوتایی اندازه‌گیری می‌شود. برای مثال، هر مراجعه‌کننده را می‌توان به صورت موفق یا ناموفق در خانواده‌درمانی نمره‌دهی کرد؛ و یا نتیجه برای هر بیمار دریافت کننده درمان سرطان را به صورت زنده مانده یا فوت شده (یعنی دوتایی) نمره‌گذاری کرد. این قبیل مطالعات را می‌توان در قالب یک جدول چهارخانه‌ای نشان داد که مداخله–کنترل به عنوان یک عامل و سطوح دوتایی متغییر نتیجه نیز به عنوان عامل دیگر در نظر گرفته می‌شود. در این موارد هم تفاضل میانگین استاندارد شده، و هم ضریب همبستگی، تخمینهای مسأله‌داری از اندازه اثر بدست خواهند داد. از این رو، شاخص اندازه اثر مناسب، معمولاً نرخ غیرعادیها خواهد بود که به صورت زیر تعریف می‌شود:

$$O_i = \frac{AD}{BC}$$

که در این مطالعه A، B، C و D فراوانی هر خانه در جدول ۱۳.۱ است. اگر مقدار یک خانه برابر صفر باشد، آنگاه مقدار ۰/۵ باید به تمامی خانه‌ها برای آن اندازه نمونه اضافه شود. نرخ غیرعادیها را می‌توان همچنین از روی اطلاعات دیگر مانند داده‌های موجود در مورد نسبت شرکت‌کنندگان موفق در هر گروه نیز محاسبه کرد. برای مثال، فرض کنید گروه مداخله یک واکسن بر علیه یک بیماری مشخص دریافت کرده و گروه کنترل دریافت نکرده باشد، نتیجه می‌تواند تعداد افرادی در هر گروه باشد که واکسن را دریافت کرده و یا نکرده‌اند. فرض کنید A برابر است با ۲۶۵ شرکت‌کننده واکسینه‌شده که مصون مانده‌اند و B برابر است با ۳۲ نفر از شرکت‌کنندگان واکسینه شده‌ای که به بیماری دچار شده‌اند. C برابر است با ۲۰۴ شرکت‌کننده‌ای که درمان را دریافت نکرده‌اند (در گروه کنترل) و مصون مانده‌اند و D برابر است با ۷۵ نفر از گروه کنترل که به بیماری دچار شده‌اند. شانس مصون ماندن نسبت به بیماری اگر واکسن دریافت کرده باشید عبارتست از

$$\frac{A}{B} = \frac{۲۶۵}{۳۲} = ۸/۲۸$$

بنابراین افراد دریافت‌کننده واکسن ۸ برابر بیشتر احتمال دارد که سالم بمانند تا مریض شوند. در مقابل، در گروه کنترل شانس برابر است با

$$\frac{C}{D} = \frac{۲۰۴}{۷۵} = ۲/۷۲$$

نرخ دو شانس نسبت به یکدیگر برابر است با $O_i = ۳/۰۴$؛ به این معنی که شانس این که افراد دریافت‌کننده واکسن نسبت به بیماری مصون بمانند، حدود ۳ برابر بیشتر از افرادی است که واکسن را نزده‌اند. به ندرت اتفاق می‌افتد که تمام مطالعات از یک مقیاس مشابه برای اندازه‌گیری نتایج استفاده کنند. برای مثال، مطالعات مربوط به مداخلات مرتبط با چاقی معمولاً همگی نتایج را در قالب پوند یا کیلوگرم گزارش می‌کنند (Haddock et al., 1994)، یا مطالعات درمان افسردگی نتایج را با آزمون بک محاسبه می‌کنند (Robinson, Berman, & Neimeyer, 1990)، یا مطالعات مداخلات پیش از جراحی ممکن است همگی تعداد بستری شدن روزهای پیش از جراحی را گزارش کنند (Mumford, Schlesinger, Glass, Patrick, & Cuerdon, 1984). در این موارد منطقیست که تفاضل میان میانگینهای خام را به عنوان مقیاس عمومی نتایج در نظر بگیریم. در هر حال، داده‌ها در یک مقیاس مشترک هستند و متغیر اولیه که به پوند، روز، یا دلار ارائه شده ممکن است بسیار قابل تفسیرتر از دیگر مقیاسهای اندازه‌گیری اندازه اثر باشد.

تحلیل داده‌های متاآنالیز

داده‌های متاآنالیزی که در قالب اندازه اثر ارائه می‌شوند، مانند هر داده اجتماعی و رفتاری دیگری تحلیل می‌شوند، یعنی با استفاده از آمار استنباطی و توصیفی و با تکنیکهای تک‌متغیره و چندمتغیره (Cooper & Hedges, 1994a; Hedges & Olkin, 1985). متاآنالیز داده‌های مبتنی بر مطالعه[۴۸۳] را با اطلاعات مبتنی بر یک مطالعه خاص جایگزین می‌نماید (اگرچه موارد معدودی در مورد داده‌های متاآنالیزی منحصر به فرد است که کمی جلوتر به آن خواهیم پرداخت). البته در سالهای اولیه معرفی روش متاآنالیز، محققین آماره‌های مرسوم در مطالعات اولیه را به عاریت می‌گرفتند. برای مثال، میانگین اندازه اثر و واریانس آن را در میان تمام مطالعات محاسبه می‌کردند. بنابراین اسمیت (Smith et al., 1980) میانگین ۱۷۶۶ تفاضل میانگین استاندارد شده را که از ۴۷۵ مطالعه کنترل‌شده رواندرمانی بدست آورده بود را محاسبه کرد (بیشتر مطالعات اندازه اثرهای متعددی تولید کرده بودند) و میانگین d= ۰/۸۵ را با خطای استاندارد ۰/۰۳ را بدست آورد. فاصله اطمینان ها نیز با به شیوه معمول یعنی با ضرب کردن خطای استاندارد در برخی مقادیر بحرانی (اغلب مقدار ۱/۹۶ که نمره آماره Z برای آزمون دو طرفه در سطح اطمینان ۵٪ است) و جمع و تفریق کردن حاصلضرب با میانگین برای بدست آوردن فاصله اطمینان (در این مورد فاصله اطمینان در سطح ۹۵٪ طیفی از ۰/۷۹ تا ۰/۹۱ را در بر می گیرد). از آنجا که صفر در فاصله اطمینان وجود ندارد، اسمیت و همکارانش چنین نتیجه گرفتند که میانگین لندازه اثر به میزان معناداری از صفر متفاوت است. به همین شیوه، محاسبه میانگین برای طبقه‌های دقیق طراحی شده از مطالعات متداول است. برای مثال، اسمیت و همکارانش (۱۹۸۰) دریافتند که متوسط اندازه اثر برای درمانهای رواندرمانی d = ۰/۷۸ است، برای مطالعات رفتاری ۰/۹۱ = d ، برای مطالعات شناختی ۱/۳۱ = d، برای مطالعات انسانشناسی ۰/۶۳ = d، برای مطالعات توسعه‌ای ۰/۴۲ = d، و برای مطالعات شناختی–رفتاری ۱/۲۴ = d . به طور کلی در متاآنالیز می‌توان فرضیاتی را در مورد انواع مداخله‌ها، شرکت‌کنندگان و روش‌ها مورد آزمون قرار داد. اندازه اثرها می‌توانند با مطالعه پیوسته یا مشخصات مداخله همبستگی داشته باشند. می‌توان برای یافتن اطنابها و اضافات و به منظور تعدیل عناصر غیرمرتبط مطالعه که تا به حال محاسبه شده‌اند، از

[483] Study-based data

آزمونهای چندمتغیره استفاده کرد. و در نهایت، برای ارزیابی نحوه اثر تعدیل کننده‌های احتمالی بر اندازه و جهت اثر می‌توان از طبقه‌بندی کردن بهره برد. اگرچه مولفه‌های متعددی در متاآنالیز وجود دارد که غیرعادی هستند:

۱) مطلوبیت وزن دهی تخمین های اندازه اثر با استفاده از تابعی از اندازه نمونه؛

۲) آزمون کردن همگنی اندازه اثر؛

۳) ماهیت سلسله مراتبی داده‌های متاآنالیزی؛

۴) وابستگی اندازه اثر در درون مطالعات، و

۵) وجود سوگیری انتشارات.

اول، مطالعات قبل از محاسبه میانگین وزن‌دهی می‌شوند که می‌تواند منجر به بدست آوردن پاسخ‌های کاملاً متفاوت (نسبت به تحلیلهایی که در آن وزندهی انجام نشده باشد) شود (حداقل مربعات معمولی). معروفترین چهارچوب، وزن دادن به اندازه اثر بر مبنای اندازه نمونه و یا قالبی از آن است، با این فرض که مطالعات با شرکت‌کنندگان بیشتر تخمین‌های درستتری از پارامترهای جمعیت بدست می‌دهد. اینگونه وزن‌دهی‌ها واریانس لندازه اثر متوسط را کاهش داده و بنابراین آزمونهای استنباطی را قویتر می‌نماید. دیگر چهارچوبها عبارتند از وزندهی بر اساس پایایی و روایی مقیاس‌های اندازه‌گیری نتایج (Hunter & Schmidt, 1994)، و یا بر اساس کیفیت مطالعه (Amato & Keith, 1991; Begg et al., 1996; Franklin, Grant, Corcoran, Miller, & Bultman, 1997; Jadad et al., 1996; Moher et al., 1995; Moher & Oklin, 1995; Wampler & Serovich, 1996; Wortman, 1994) انجام می‌دهند. اگرچه در کل تنها چهارچوب‌های وزن‌دهی بر مبنای اندازه نمونه و شاخصهای روانسنجی به لحاظ منطقی و آماری به خوبی تبیین شده‌اند، بررسی و اندازه‌گیری متغیرهایی مانند کیفیت مطالعه احتمالاً در طول تجزیه و تحلیلهای بعدی روی داده‌ها بهتر بدست می‌آید تا در مرحله وزن‌دهی (Shadish & Haddock, 1994).

دومین مولفه تحلیل داده‌های متاآنالیزی که می‌تواند غیرمعمول بنظر برسد، آزمون همگنی یا تجانس است. با استفاده از این آزمون می‌توان این موضوع را بررسی کرد که آیا میزان واریانس مجموعه اندازه اثرهای مشاهده شده تنها به میزانی است می توان به خطای نمونه‌گیری نسبت داد یا نه (بخشی از تفاوت میان اندازه اثر جامعه و اندازه اثر نمونه به دلیل این واقعیت رخ می‌دهد که تنها نمونه‌ای از افراد جامعه مشاهده شده‌اند). اگر چنین است می‌توان گفت که اندازه اثرهای مشاهده‌شده همگن هستند و واریانس موجود تصادفی و غیرقابل پیش‌بینی بوده است. اگر فرض همگنی رد شود، به این معناست که توزیع اندازه اثرها شامل میزان واریانسی بیشتر از آنچه انتظار می رفت بواسطه شانس رخ دهد، بوده است (با در نظر گرفتن این که اندازه نمونه‌ها در مطالعات تحت بررسی تجمیع شده‌اند). همچنین ممکن است بتوان واریانس احتمالی را با استفاده از دیگر متغیرهای نظری یا روانشناختی پیش‌بینی کرد. برای مثال لیپسی (Lipsey, 1992) اثر مداخلات مرتبط با بزهکاری نوجوانان را در ۳۹۷ مطالعه مورد بررسی قرار داد و یک متوسط وزنی از اندازه اثر برابر با ۰/۱۰۳ = d بدست آورد. اگرچه فرض همگنی در مورد اندازه اثرهای بدست آمده رد شد، زیرا اندازه اثرها سه برابر بیشتر از میزان مورد انتظار بواسطه شانس

واریانس داشته‌اند. بنابراین لیپسی مدلهای رگرسیونی متنوعی را برای یافتن متغیرهایی که مسئول ایجاد این واریانس هستند بکار گرفت. یقیناً رد همگنی در یک مورد تک متغیره تعجب برانگیز نیست. زیرا همگنی در این زمینه نشان‌دهنده این است که یک و تنها یک متغیر (دریافت مداخله مرتبط با بزهکاری نوجوانان از هر نوع) مسئول تمامی واریانس سیستماتیک در نتایج این مداخلات است. البته ندرتاً انتظار می رود که چنین چیزی صحت داشته باشد. بلکه در علوم اجتماعی به طور معمول انتظار داریم چندین متغیر پدیده مورد بررسی ما را تبیین کنند و این استراتژی در متاآنالیز نیز موفقیت آمیز است (Shadish, 1992a).

برخی مواقع همگنی با وجود در نظر گرفتن تمامی متغیرهای در دسترس باز رد می‌شود (Lipsey, 1992). در این مواقع دو امکان وجود دارد. یکی اینکه متغیرهای دیگری وجود دارند که در صورت کد شدن آنها همگنی بدست خواهد آمد. امکان دوم آن است که جمعیتی که اندازه اثرها از آن گرفته شده به جای یک اندازه اثر (مدل اثرات ثابت) دارای توزیعی از اندازه اثرها (مدل اثرات تصادفی) هستند. اگر چنین حالتی صادق باشد اضافه کردن متغیرهای بیشتر ما را به همگنی نمی‌رساند، بلکه تخمین زننده‌های اثرات تصادفی می‌توانند برای لحاظ کردن این امکان محاسبه شوند (Hedges & Vevea, 1998; Louis & Zelterman, 1994; Raudenbush, 1994; Shadish & Haddock, 1994). یقیناً تحلیل‌گران بیضی از مدل‌های اثرات تصادفی به عنوان یک اصل پشتیبانی می‌کنند؛ فارغ از اینکه نتایج آزمون همگنی چه باشد. به زعم آنها استفاده از این مدلها ماهیت نامطمئن استنباطها در تمامی انواع داده‌ها (که داده‌های متاآنالیزی را نیز شامل می‌شود) را نشان می‌دهد (Raudenbush, 1994). مدلهای اثرات تصادفی غالباً خطای استاندارد و فاصله اطمینان را افزایش می‌دهند. این افزایش نشان‌دهنده عدم‌قطعیت بیشتر نسبت به استنباطهای انجام‌شده در مورد جمعیت است.

سومین دلیل غیرمعمول بودن داده‌های متاآنالیزی آن است که پاسخ‌دهندگان درون مطالعات لانه‌گزینی کرده‌اند. بحث مربوط به داده‌های لانه‌گزیده پیش از این مطرح شد، آنجا که در آزمایشهای تصادفی دانش‌آموزان در درون کلاسها لانه‌گزینی کرده بودند. بنابراین مشکل موسوم به مسأله واحد تحلیل به طور معمول در متاآنالیز بروز می‌کند. بهترین تحلیل‌ها برای لحاظ کردن و در نظر گرفتن مسأله لانه‌گزینی بر روشهای چند–سطحی استوار هستند (Bryk & Raudenbush, 1992; Kalaian & Raudenbush, 1996). این مدلها، مدلهای اثر تصادفی هستند به همان ترتیبی که استفاده از آنها مشکلات ناشی از غیرهمگنی را نیز مرتفع می‌سازد. گلداشتاین و همکارانش (Goldstein et al., 2000) مدل چند–سطحی ای را ارائه می‌کنند که می‌تواند برای ترکیب داده‌های در سطح متاآنالیز با داده‌های منفرد بکار گرفته شود. این ترکیب زمانی به کار می‌آید که برخی مطالعات آمارهای خلاصه را (خلاصه مطالعات) گزارش می‌کنند، در حالی که در برخی دیگر از مطالعات داده‌های منفرد گزارش شده است.

چهارمین مشخصه متفاوت مطالعات متاآنالیزی آن است که هنگامی که مطالعه‌ای چندین اندازه اثر داشته باشد، این اندازه‌ها به صورت کاتوره‌ای (استاکستیک)[۴۸۴] وابسته هستند. این موقعیت زمانی می‌تواند بروز کند که مطالعه مقیاس‌های نتیجه‌ای متعدد داشته باشد، مطالعه نتایج را در مورد آن متغیرهای متعدد چندین بار در طول زمان گزارش می‌کند، یا مطالعه مداخله‌های متعددی را با یک گروه کنترل مشترک در مورد تنها یک متغیر میانجی مقایسه می‌کند. این وابستگی‌ها

[484] stochastic

نقض‌کننده فرض استقلال اندازه اثرهاست. بنابراین متداول است که قبل از ترکیب مطالعه‌ها، مقدار متوسط اندازه اثرها از مقیاسهای مختلف درون یک مطالعه محاسبه شود. انجام این کار به معنی آن است که اندازه اثرها کاملاً همبستگی دارند. اگر همبستگی میان مقیاسها شناخته شده باشد، می‌توان با استفاده از رویه‌های چندمتغیره خاصی تخمین‌های میانگین اندازه اثر کارآمدتری (با تغییرات محدودتر) بدست آورد (Gleser & Olkin, 1994; Hedges & Olkin, in press; Kalaian & Raudenbush, 1996; Raudenbush, Becker, & Kalaian, 1988). هنگامی که همبستگی میان مقیاسها بالاست، فایده چندانی از نظر کارایی حاصل نمی‌شود؛ زمانی که همبستگی به صفر می‌رسد، فایده از نظر افزایش کارآمدی می‌تواند میان ۱۰ تا ۱۵٪ باشد. اما هنگامی که همبستگی پایین باشد، مشخص نیست که متغیرهای وابسته مختلف باید ترکیب شوند یا به صورت مجزا تحلیل شوند (Harris & Rosenthal, 1985).

پنجمین مشخصه غیرمعمول، وجود سوگیری انتشارات است. این مسأله از آن جهت بوجود می‌آید که بسیاری از داوران مجلات، از چاپ مقالات با یافته‌های مرتبط با اثرات غیرمعنادار امتناع می‌نمایند. بنابراین، چنین مقالاتی چاپ نشده رها می‌شوند و بسیار دشوار خواهد بود که آنها را پیدا کرده و برای مقایسه کردن با مطالعات با یافته‌هایی در مورد اثرات معنادار در متاآنالیز لحاظ نمود. و از آنجا که معناداری و اندازه اثر همبستگی دارند، مطالعات به چاپ رسیده غالباً اندازه اثر را بیشتر از حد واقعی تخمین می‌زنند. یک راه غلب برای حل این مشکل، محاسبه تعداد مطالعات غیرمعناداری است که برای غیرمعنادار شدن یک اندازه اثر متوسط مورد نیاز است (Orwin, 1983; Rosenthal, 1979; Iyengar & Greenhouse, 1988; Schonemann, 1991). البته تکنیکهای پیچیده‌تری نیز برای مواجهه با سوگیری انتشارات وجود دارد (Begg, 1994; Duval & Tweedie, 2000; Greenhouse & Iyengar, 1994; Hedges, 1992; Hedges & Olkin, in press; Hedges & Vevea, 1996; Silliman, 1997).

تفسیر و گزارش نتایج

با وجود ادبیات تخصصی در حوزه تحقیق، به طور کلی تفسیر و ارائه نتایج متاآنالیز با مشکلات خاصی همراه است (Halvorsen, 1994; Light, Singer, & Willett, 1994; Wang & Bushman, 1998). تفسیرها غالباً به طور اجتناب ناپذیری به خارج از محدوده داده‌های بررسی شده کشیده می‌شوند. محقق معمولاً با توجه به گرایش‌های نظری خاصی که دارد، معانی خاصی را به داده‌ها نسبت می‌دهد که بعضاً ناصحیح بوده و خالی از سوگیری نیست. در هنگام تفسیر نتایج متاآنالیز خصوصاً برای اهداف مرتبط با استنباطهای علّی، یک نکته اساسی باید همواره در نظر باشد و آن اینکه داده‌های حاصل از متاآنالیز از نوع همبستگی هستند (Louis, Fineberg, & Mosteller, 1985; Shadish, 1992a). محققین متاآنالیز هیچگاه مطالعات را به طور تصادفی به افراد، موقعیتها، زمانها، علّتها و اندازه اثرهای تحت بررسی تخصیص نمی‌دهند. این خود فی‌النفسه دشوار است که بتوان تصور کرد انجام چنین تخصیصی چطور می‌تواند ممکن باشد. به همین ترتیب به ندرت در متاآنالیز عناصر طرح شبه‌آزمایشی که برای بی‌اثر کردن تهدیدات روایی بکار گرفت می‌شود، را مورد استفاده قرار می‌دهیم. تقریباً در تمامی موارد، محقق طی انجام متاآنالیز، داده‌های مشاهده‌ای را به روشی کاملاً شبیه به فرایند یک پیمایش ثبت می‌کند؛ که این موضوع باعث می‌شود تا استنباطهای علّی صورت گرفته در مورد اثرات ناشی از عضویت در طبقات اندازه اثر، مبتلا به همان تهدیدات روایی باشد که در مورد هر داده همبستگی دیگری وجود دارد. دسته‌ها و طبقه‌های مورد توجه

علّی در یک متاآنالیز عموماً با بسیاری از دیگر طبقه‌های ناشناخته یا محاسبه نشده در متاآنالیز مخلوط[۴۸۵] هستند. شبیه همان چیزی که در مطالعات همبستگی یا شبه‌آزمایشی دیده می‌شود (Knight, Fabes, & Higgins, 1996).

تنها اســتثناء زمانی رخ می‌دهد که مجموعه‌ای از آزمایشــهای تصــادفی در مورد یک مداخله خاص را مورد بررســی قرار می‌دهیم. در این حالت، اســتنباط علّی متاآنالیزی در مورد اثر مداخله در میان تمامی مطالعات (لحاظ شــده در متاآنالیز) ضــعیف‌تر از اســتنباطهای بدســت آمده از هر کدام از مطالعات (به طور منفرد) نخواهد بود. برای مثال، متاآنالیز ۷۱ آزمایش تصــادفی در مورد اثرات رواندرمانی خانواده در مقایســه با کنترلهای تصــادفی به مقدار d=۰/۵۱ در پس–آزمون رســید (Shadish et al., 1993). می‌توان چنین نتیجه‌گیری کرد که این رواندرمانیها باعث بوجود آمدن چنین اندازه اثرهایی شــده اســت (حداقل با همان قوتی که در مطالعات اولیه ادعا کردیم). اگرچه این نتیجه‌گیری تنها در مورد یک اســتنباط منفرد در خصــوص اثر یک مداخله مشــترک قابل انجام اســت. اســتنباط در مورد دیگر طبقه‌ها مانند اثربخشــی مقایســه‌ای درمانهای مختلف زیرمجموعه یک درمان که در مطالعات مختلف و یا با جزئیات روشهای شناختی مختلف مورد بررسی قرار گرفته‌اند، همچنان مطالعه همبستگی محسوب می‌شود.

متاانالیز و پنج اصل استنباط علّی تعمیم یافته

متاآنالیز کاربردهای فراوانی برای تعمیم نتایج دارد. در کل، محاســبه متوســط یا میانگین چند مطالعه ابزاری برای توضــیح جهت‌گیری کلی آن مطالعه اســت. از جمله نقدهای اولیه بر متاآنالیز این تعیم‌ها را به درســتی به چالش می‌کشــد. زیرا اندازه اثرها، حاصــل قضــاوتهای متنوع و متعددی از ســوی محققین مجری مطالعات اولیه و محقق مجری متاآنالیز خواهد بود. این قضــاوتها شــامل قضــاوت درباره اینکه آیا اصــلاً یک مطالعه را می‌باید انجام داد یا نه نیز صــورت می‌گیرد؛ و اینکه آیا آن را بنویســند یا نه، چه چیزهایی را در جریان نگارش گزارش کنند، چطور مطالعه را طراحی کنند، آیا نتایج را بایگانی کنند، و آیا یک مطالعه را در متاآنالیز لحاظ کنند (با توجه به روش‌شناسی و عنوان آن). در بهترین حالت، در میان این قضاوتها، میانگین گرفتن باعث از دســت دادن مقداری از اطلاعات می‌شــود، و در بدترین حالت، باعث پنهان ماندن واریانســهای با اهمیت می‌شود. نقدها البته به مسائل عمیقتر نیز کشیده شده است. مثلاً اینکه برای یک سوال خاص، اندازه اثرهای مورد بررسی به صــورت تصــادفی از جمعیت تمامی اندازه اثرهای ممکن گرفته نمی‌شــود. از این بابت اســت که متاآنالیز بر یافتن تمامی مطالعات انجام شده در یک زمینه تاکید دارد (ن.ک. Rubin, 1990, 1992). اما حتی اگر تمامی مطالعات منتشر شده و نشده در مورد موضــوعی را بتوان پیدا کرد، این همچنان نشــان‌دهنده اجماعی از تمامی مطالعات انجام شــده اســت و نه نماینده جمعیت ممکن.

علاوه بر این، حتی اگر تمامی مطالعات منتشــر شــده و نشــده مورد بررسی قرار بگیرند، و همگی دارای نوعی سوگیری ثابت باشــند، متآنالیز ارزش چندانی نخواهد داشت. برای مثال، در متاآنالیز انجام شــده در مورد نحوه اثر منع جداسازی کودکان در مدارس بر دستاوردهای تحصیلی کودکان سیاه پوست، چنین مسأله‌ای رخ داد. اگرچه کودکان برای ۱۰ سال یا بیشتر در مدرسه حضور داشتند، مطالعات موجود تنها ۲ سال اول بعد از منع جداسازی را پوشش می‌دادند. که این احتمالاً منجر می‌شد

[485] Confound

تا اثر منع جداسازی کمتر از مقدار واقعی تخمین زده شود. واقعیت آن است که متاآنالیز با نمونه‌های هدفمند اما غیرهمگن از مطالعات کار می‌کند، جمعیتی مورد نمایندگی آنها به ندرت روشن بوده و الزاماً مبرا از سوگیریهای پایدار و ثابت نیستند. با این همه، این نمونه‌های اخذ شده از مطالعات همچنان بی‌اندازه برای استنباطهای علّی تعمیم‌یافته سودمند هستند. در ادامه این فصل، با بکارگیری پنج اصل استنباط علّی تعمیم‌یافته نشان خواهیم داد که علّت این امر چیست.

شباهت سطح

در این اصل، تاکید بر ارزیابی تناسب میان عملیات تحقیق و عناصر نوعی[۴۸۶] اهداف مورد تعمیم است. متاآنالیزهای انجام شده روی مطالعات متعدد به دو صورت می‌توانند این امر را تسهیل کنند. اول، مطالعات متعدد تعداد بیشتری سازه را نشان می‌دهند (در مقایسه با تعداد سازه‌ای که یک مطالعه می‌تواند نشان دهد). بنابراین تعداد استنباطهای قابل انجام را افزایش می‌دهند. به دلایل منطقی، غالب مطالعات انجام شده روی برنامه‌های آموزش بیماران، خود را به نوع خاصی از بیماران یا حداکثر انواع محدودی از بیماران محدود می‌کنند. اما در طول مطالعات متعدد، انواع متنوعی از بیماران مورد بررسی و مطالعه قرار گرفته‌اند (برای مثال بیماران زنان، اورولوژی، ارتوپدی ...). به همین ترتیب، تقریباً تمام مطالعات در یک شرایط آزمایش خاص صورت می‌پذیرند، مثلاً یک بیمارستان خصوصی بزرگ در یک شهر اصلی و بزرگ، اما متاآنالیز مطالعات می‌تواند در مورد شرایط مختلف مطالعاتی که در بیمارستانهای بزرگ، کوچک، خصوصی، دولتی، شهری یا حومه‌ای و حتی بیمارستانی در مقایسه با درمانگاه نتیجه‌گیری‌هایی ارائه نماید (Devine, 1992).

دوم، متاآنالیزها این شانس را دارند که مطالعاتی را بیابند که در آن از عملیاتی بهره برده شده که سازه‌های خاص را نشان می‌دهد. این مسأله خصوصاً زمانی که سازه نایاب است اهمیت پیدا می‌کند. برای مثال، برخی مکاتب رواندرمانی از شیوه‌ای نسبتاً نایاب برای درمان استفاده می‌کنند که به آن مداخلات متناقض گفته می‌شود. در این مداخلات، درمانگر به بیماران پیشنهاداتی ارائه می‌دهد که با منطق سازگار نیست و کمتر احتمال می‌رود موجب بهبود بیمار شود. شوهام و همکارش (Shoham-Salomon & Rosenthal, 1987) بوسیله متاآنالیز صدها مطالعه رواندرمانی توانستند ده مطالعه که در آنها این روش کمیاب بکار گرفته شده بود را پیدا کنند. با این وجود، برخی مواقع هیچ مطالعه‌ای در مورد موضوع تعمیم وجود ندارد. در این موارد، یکی از مزایای اصلی متاآنالیزها آن است که سازه‌های مهمی که تا به حال مغفول واقع شده‌اند را نمایان ساخته، و از این طریق راه را برای مطالعه آنها در تحقیقات بعدی هموار می‌سازند. به عنوان نمونه، حتی کسانی که با اثر رواندرمانی در شرایط کلینیکی مخالف هستند، بر این مسأله متفق‌القول هستند که مطالعات بسیار معدودی در این زمینه وجود دارد که هم از طراحی مناسبی برخوردار بوده، و هم تمامی عناصر نوعی این قبیل درمانهای در آن لحاظ شده باشد. با توجه به این مساله، آژانسهای سرمایه‌گذاری در آمریکا مانند موسسه ملی سلامت روان بودجه‌های تحقیقاتی مشخصی را برای انجام اینگونه مطالعات اختصاص داده‌اند (Shadish et al., 1997; Shadish et al., 2000; Weisz et al., 1992).

ضرورتی ندارد محقق دانش کاملی نسبت به مشخصات نوعی سازه موردنظر، ورای آنچه در مطالعات تحت بررسی در متاآنالیز انجام شده است، داشته باشد. استنباطهای مرتبط با سازه‌ها را می‌توان بر مبنای عملیات صورت گرفته در مطالعات

[486] prototypical

متعدد و دانش موجود بدست آورد. اگرچه داشتن دانش نسبت به سازه می‌تواند به عنوان منبعی برای ارزیابی روایی نتایج بدست آمده از متاآنالیز بکار بیاید. برخی مواقع داده‌هایی درباره موضوع تعمیم وجود دارد، و بنابراین می‌توان عملیات نمونه را با آن داده‌ها مقایسه نمود. برای مثال، متاآنالیز انجام شده توسط شدیش و همکارانش (Shadish et al., 1997) روی رولندرمانی درمانهای کلینیکی ادعاهای اولیه در مورد تجربیات رولندرمانی مودال[۴۸۷] را با ارجاع به داده‌های تجربی که پیش از آن برای تشریح چنین درمانهایی جمع‌آوری شده بودند، پشتیبانی نمود. در برخی دیگر از مواقع، دانش تنها با استفاده از خرد حاصل می شود. برای مثال، میر و همکارش (Meyer & Mark, 1995) متوجه شدند که تقریباً تمامی ۴۵ مطالعه ای مورد متاانالیز در خصوص مداخلات اجتماعی–روانی روی افراد بالغ مبتلا به سرطان، روی زنان سفید پوست آمریکایی صورت گرفته بودند. این گروه به طور قطع زیر–مجموعه ی کوچکی از تمامی افراد بالغ مبتلا به سرطان به حساب می‌آیند. در نهایت، آنچه به عنوان مشخصات نوعی یک سازه در نظر گرفته می‌شود، انعکاس‌دهنده درک مشترکی است که نسبت به آن سازه در میان افراد و سازمانهای متخصص در آن موضوع رواج دارد. عناصری که جامعه علمی آنها را به عنوان «مشخصه نوعی» برای یک سازه در نظر می‌گیرد، در بیشتر نمونه‌های مطالعات یافت خواهند شد و بالعکس عناصری که کمتر به عنوان نوعی برای یک سازه شناخته می‌شوند کمتر در مطالعات در نظر گرفته می‌شوند. برای مثال، جامعه علمی در مورد سازه افسردگی متفق‌القول است که این سازه دارای سه عنصر شناختی، احساسی و روانی است (Tanaka & Huba, 1984). از آنجا که مقیاس بِک (Beck Depression Inventory; Beck et la., 1996) این عناصر و ابعاد را ارزیابی می‌کند این مقیاس به طور وسیعی مورد استفاده قرار گرفته است. در نتیجه متاآنالیز مطالعات انجام شده در زمینه افسردگی، انعکاس‌دهنده اجماع موجود در خصوص سازه افسردگی خواهد بود.

یقیناً تحقیقات بعدی در آینده خواهند توانست سوگیریها و خطاهای سازه‌ها را بررسی کرده و نشان دهند. این سوگیریها با گذشت زمان و کسب دانش بیشتر در مورد سازه قابل تشخیص خواهند شد. متاآنالیزها بهتر از مطالعات منفرد قادر به نشان دادن سوگیریها خواهند بود. اما آنچه اهمیت دارد آن است که در نظر داشته باشیم هدف دستاوردهای تحقیق نیست، بلکه هدف یافتن سوگیریهای پایدار و ثابتی است که بواسطه فرایندهای نمونه‌گیری مجال بروز می‌یابند، حتی زمانی که مطالعات متعددی مورد متاآنالیز واقع می‌شوند.

بی‌اثر کردن غیرمرتبطها[۴۸۸]

اصل شباهت سطح فی‌النفسه نمی‌تواند سوگیری‌ها را آشکار سازد. باید بتوانیم انحرافهای موجود نسبت به درک عمومی از سازه را به دو دسته‌ی مواردی که موجب تغییر در استنباط آماری می‌شوند، و مواردی که این اثر را ندارند تفکیک نماییم. انحرافاتی که باعث تغییر نتایج استنباطهای علّی نمی‌شوند غیرمرتبط هستند، و می‌توانیم از آنها میانگین بگیریم بدون آنکه تعمیم[۴۸۹] در نتایج را از دست بدهیم. اصل بی‌اثر کردن غیرمرتبطها، سعی در تفکیک انحرافات غیرمرتبطها از انحرافات مرتبط دارد. اینکار از طریق تأکید بر گفتمان رایج میان اعضای یک جامعه علمی در خصوص متغیرهای تعدیلگر برای یک

[487] modal
[488] irrelevancies
[489] generality

رابطه علّی، و همچنین استفاده از تحلیلهای اکتشافی برای یافتن انحرافاتی که خود موجب ایجاد انحراف دیگر می‌شوند صورت می پذیرد. هیچکدام از این دو رویه – بررسی گفتمان غالب یا مطالعات اکتشافی- کامل و بی‌نقص نیستند؛ و برخی سوگیریهای پایدار تنها در نقدهای وارد آمده از سوی افراد خارج از جامعه محققین آن زمینه تحقیقاتی و در طول دوره‌های زمانی طولانی شناسایی می‌شوند. با این وجود، جستجوی جدی برای متغیرهای تعدیلگر کمک می‌کند منابع ایجاد کننده ناهمگونی مرتبط و غیرمرتبط را شناسایی کنیم. زیرا این ناهمگونیها می‌توانند تعمیمهای علّی را محدود نمایند.

بحث را با مثال آموزش بیماران ادامه می‌دهیم. یک باور مشترک وجود دارد که اینگونه آموزشها اثر یکسانی دارند، فارغ از اینکه چه نوع بیمارستانی مورد مطالعه قرار گرفته است. بر این مبنا، اگر تامین‌کنندگان پروژه‌ها و داوران مقالات اهمیت اندکی برای نوع بیمارستان قائل باشند، و محققین بیمارستانها را به صورت کاتوره‌ای و شانسی انتخاب کنند، بیمارستانهای انتخاب شده برای متاآنالیز از تمامی مناطق کشور خواهند بود. اصل بی‌اثر کردن غیرمرتبطها باعث می‌شود تا متاآنالیز نشان دهد که رابطه علّی موردنظر با حضور تمامی این عناصر غیرمرتبط وجود دارد. و از آنجا که این ناهمگونیها در مطالعات متاآنالیز بیشتر از هر یک از مطالعات منفرد وجود دارد، پتانسیل و توان متاآنالیز برای ارائه تعمیم‌های علّی بیشتر از هر یک از مطالعات منفرد خواهد بود.

هنگامی که تعداد و طیف متغیرهای مرتبط با شرایط آزمایش‌که به طور شهودی انتخاب می‌شوند زیادست، یک اندازه اثر متوسط را می‌توان از روی تمامی بیمارستانهای موجود اندازه‌گیری کرد و با استفاده از آن این مسأله را ارزیابی کرد که آیا یک رابطه علّی به میزان کافی مستحکم است که بتواند یا وجود واریانسهای موجود همچنان پایداری نماید؟ اینگونه آنالیزها احتمال آن را که اثر مشاهده شده با اثر نوع بیمارستان مخلوط نشده باشد را افزایش می‌دهد. از این رو، نشان می‌دهد که آیا رابطه علّی موردنظر علیرغم واریانس قابل‌توجه در نوع بیمارستانها، در متاآنالیز حاضر همچنان معنادار است؟ اما بهتر است به جای تجمیع بیمارستانها با یکدیگر در قالب مجموعه‌ای ناهمگون از بیمارستانها، نوعی گونه‌شناسی اختصاصی از بیمارستانها ایجاد کنیم و این گونه‌شناسی را در تحلیل داده‌ها وارد نماییم. در این حالت می‌توانیم بررسی کنیم که آیا رابطه علّی برای انواع مختلف بیمارستانها، مناطق، بیماران، و یا شیوه‌های مختلف مفهوم‌پردازی علّت همچنان صادق است یا نه. مطالعه دوین (Devine, 1992) مصداق بکارگیری این روش است. این مطالعه نشان داد که آموزش بیمار، فارغ از نوع بیماری، نوع بیمارستان، دوره‌های زمانی، و شیوه‌های مختلف مفهوم‌پردازی آموزش بیماران، بهبود، بازتوانی و بهبود از جراحی را تسریع می‌بخشد. چنین اثربخشی محکمی در نمونه‌ای از بیش از ۱۵۰ مطالعه این نتیجه‌گیری را که آموزش بیماران احتمالاً اثرات قابل‌تعمیم قابل‌توجهی دارد را تقویت می‌کند. اگرچه، در صورتی که اندازه نمونه برای هر خوشه یا دسته کوچک است باید جانب احتیاط را رعایت کنیم. زیرا قدرت اندک می‌تواند باعث شود تا متاآنالیز عنصری را بی‌ارتباط قلمداد کند در حالی که، در حقیقت، متاآنالیز نمونه بسیار ناکافی و کوچکی برای ارزیابی درست آن فرضیه داشته است.

وجود تعداد زیاد و طیف وسیعی از عناصر غیرمرتبط می‌تواند استنباطهای مرتبط با سازه‌های علّت و معلولی را در متاآنالیز ارتقاء بخشد. در متاانالیز یک علّت واحد از طریق روشهای بسیار بیشتری مورد دستکاری قرار می‌گیرد و یک اثر واحد با روشهای بسیار متنوع‌تری (در مقایسه با یک مطالعه واحد و عناصر غیرمرتبط همراه با آن مطالعه) مورد اندازه‌گیری قرار می‌گیرد. هرگونه یافته علّی بدست‌آمده با وجود این میزان از تنوع و ناهمگونی از نظر مفهومی پایه‌دارتر است؛ به ویژه

هنگامی که هر کدام از عناصر غیرمرتبط به عنوان متغیر خوشه‌بندی بکار گرفته شده باشند و رابطه علّی در شرایط وجود عنصر غیرمرتبط از میان نرفته باشد.

بعضاُ در مواردی هیچ واریانسی در سازه‌ای که غیرمرتبط فرض می‌شود وجود ندارد، بنابراین بررسی مرتبط یا غیرمرتبط بودن آن سازه عملاُ ناممکن خواهد بود. برای مثال، مطالعه دوین و کوک (Devine & Cook, 1983, 1986) نشان داد که تقریباً در تمامی مطالعات موجود در زمینه آموزش بیماران، از محققین به جای پرستاران برای انجام مداخلات استفاده شده بود. در حالی که پرستاران قاعدتاً به اندازه محققین در اجرای آموزش به بیماران متبحر نخواهند بود، و بعلاوه مسئولیتهای زمانبر دیگری نیز در بیمارستان بر دوش آنهاست. آیا پرستاران می‌توانند آموزشها را تا اندازه که اثربخشی مورد نظر را در پی داشته باشد، ارائه کنند؟ در زمان مطالعه دوین و کوک تنها ۴ از ۱۰۲ مطالعه‌ای که تا آن زمان در دسترس قرار داشت از پرستاران برای اجرای مداخله استفاده کرده بودند؛ و اندازه اثر بدست آمده برای این چهار مطالعه کمتر از مطالعاتی بود که در آن محققین مداخلات را اجرا کرده بودند. اگرچه مداخلات صورت گرفته در این چهار مطالعه از جامعیت کمتری نسبت به پروتکل بهینه آموزش بیماران برخوردار بود، و مطالعه بعدی که در آن پرستاران معمولی مداخلات قویتری را اجرا کردند منجر به کسب نتایجی شبیه نتایج بدست آمده در متاانالیز شد (Devine et al., 1990). با این وجود، این واقعیت که نوع افراد اجرا کننده مداخلات به طور سیستماتیک با قدرت اثر یافت شده در ۱۰۲ مطالعه همبستگی دارد، نشان می‌دهد که تعداد مطالعات لحاظ شده در متاآنالیز نسبت به ارتباط آن مطالعات به مسائل مفهومی، از اهمیت کمتری برخوردار است.

دیرکتور (Director, 1979) یافته‌های مشابهی را در مورد سازه آموزشهای ضمن خدمت ارائه داد. وی بر این باور است که با توجه به اینکه آموزشها غالباً به افراد با سطح تحصیلات و سابقه کاری پایینتر ارائه می‌شود، نتایج دارای نوعی سوگیری است که با هیچ وسیله‌ای قابل کنترل و حذف شدن نیست.

بدیهیست که برخی متاانالیزها سعی در کاهش و نه افزایش ناهمگونی عناصر غیرمرتبط داشته‌اند. توماس چالمرز و همکارانش (Chalmers et al., 1988) به صورت روتین متاانالیزهایی روی نتایج آزمایشهای کلینیکی صورت گرفته روی داروها و رویه‌های جراحی خاص انجام می‌دادند؛ ایان چالمرز و همکارانش (Chalmers, Enkin, & Keirse, 1989) همین کار را در مورد آزمایشات تصادفی مربوط به بارداری و زایمان انجام دادند. متاانالیز آنها با دقت به دنبال یافتن مطالعاتی بود که متغیرهای مستقل، مقیاسهای اندازه‌گیری نتایج، طرحهای آزمایش، آسیب‌شناسی‌ها و متخصصین درمانگر مشابهی را مورد استفاده قرار داده بودند. مدل مطالعه آنها شبیه یک تکرار[۴۹۰] ایده‌ال دقیق است که با هدف یافتن تصویری واضح از اثربخشی علّی یک مداخله تحت شرایط کاملاً استاندارد و کنترل شده، و به منظور یافتن یک اثر خاص (در زمانی که اثر واقعا وجود داشته باشد)، انجام می‌شود.

این حالت را با دیگر حالتهای موجود در متاانالیز مقایسه کنید که در آن، هدف ارزیابی اثربخشی مداخلات تحت شرایط دنیای واقعی است، و در آنها وجود نگرشی متفاوت نسبت به منابع ناهمگونی معمول است. در مطالعه مربوط به آموزش بیماران علائمی که مورد متاانالیز واقع شدند، متنوعتر بودند. متغیرهای ارزیابی نتایج از یک مطالعه به مطالعه دیگر تغییر می‌کردند (مثلاً مدت زمان اقامت در بیمارستان، مقدار داروهای ضددرد گرفته شده، میزان رضایت از بیمارستان)، و مداخلات

[490] Replication

از مطالعه‌ای به مطالعه‌ی دیگر بسته به سطح تحصیلات بیماران و زمان و منابع در دسترس برای انجام مداخلات، متفاوت بود. در آن متاانالیز هیچ مطلعه تکرار که دقیقاً مطالعات قبل از خود را تکرار کرده باشد یافت نشد. یقیناً ناهمگونی تا آنجا که شبیه به ناهمگونیهای موجود در زمینه اجتماعی مطلوب برای اجرای مداخله باشد، بلامانع است. علاوه بر این، هرچه اثرات با وجود ناهمگونی در افراد، زمینه‌های مطالعات، درمانها، معیارهای اندازه‌گیری، و تعداد دفعات آزمایش، پایدارتر و مستحکمتر باشند، اطمینان بیشتری می‌تواند نسبت به وجود آنها در زمینه‌هایی که هنوز مورد بررسی قرار نگرفته است، وجود داشته باشد. متاانالیزهای صورت گرفته روی یک نمونه از مطالعات کاملاً استاندارد نمی‌تواند چنین نتایجی بدست دهد.

ایجاد تمایز[۴۹۱]

در اینجا استراتژی این است که نشان دهیم یک استنباط خاص، تنها برای سازه‌ی با تعریف مورد نظر مصداق داشته، و برای دیگر سازه‌های جایگزین یا تعریفهای دیگر از همان سازه معنادار نیست. برخی مواقع این تمایزات در خدمت همان اهدافی قرار می‌گیرد که در نظریه اندازه‌گیری مدنظر هستند. به این معنی که از تمایزات برای یافتن تفاوت میان ویرایش‌های متعدد یک سازه، و به منظور روشن ساختن مباحث نظری و سیاستگذاری استفاده می‌شود. برای مثال، شدیش و همکارانش (Shadish et al., 1997) سازه روان‌درمانی کلینیکی نماینده را به تآسی از مفهوم‌سازی وایزتس و همکارانش (Weisz et al., 1992) از مفهوم رواندرمانی کلنیکی مورد بررسی قرار دادند. اما در طول مطالعات بعدی خود متوجه شدند که این دو سازه به ظاهر مشابه یکسان نیستند (رواندرمانی کلینیکی طبق تعریف وایزتس پیش از شروع آزمایش در کلینیک وجود دارد اما این مشخصه در مطالعه وایزتس و همکارانش در نظر گرفته نشده است). اگرچه بررسی داده های مطالعه ی شدیش و همکارانش نشان می‌دهد که تعدادی از اعضای نمونه مورد بررسی ایشان تعریف وایزتس از سازه را مبنای کار خود قرار داده‌اند. بنابراین یکی از کارکردهای متاآنالیز این است که بحث‌های موجود درباره اینگونه سازه‌های مهم و اثرگذار در ادبیات را تبیین و تنویر نماید، تا محققین بتوانند به نحو بهتری مشخصات سازه‌های مطالعه شده یا نشده در ادبیات را درک نمایند. مهمتر از آن، تبیین اینگونه تمایزات در ساختار معنایی سازه‌ها در تعیین و تحدید مرزهای استنباطهای علّی سودمند است. برای مثال، در تحقیقات رواندرمانی برخی محتوی کلامی پیام درمانگر و یا تکلیفی که وی برای بیمار در کنترل و پایش رفتارهای خود تعیین می‌کند را به عنوان تعریف سازه مداخله یا مداخله در نظر می‌گیرند. اما برخی دیگر، از محققین مداخله را بیشتر در قالب دارونماها تعریف می‌کنند و وجود رابطه علّی را در اثر پرداخت حق ویزیت به رواندرمانگر و یا هر فرد دیگری که دارای مدارک رسمی برای کمک روانی به دیگران است، می‌دانند. در نتیجه، رواندرمانگران در تلاشند تا اثر رواندرمانی را از اثر درونماها تمیز دهند. در پزشکی، قدرت دارونما تا به آن حد پذیرفته شده است که مطالعات کور–دوباره[۴۹۲] برای بی‌اثر کردن احتمال وجود این‌گونه تعابیر توصیه می‌شود. اما دارونما تنها تهدید موجود برای تعمیم اثر رواندرمانی نیست. برخی مواقع به دلیل بازتعریف محتویات فعال مداخله چالشهایی بوجود می‌آید. فرض کنید محتویات درمان، محتوای کلامی و یا فعالیتهای تعیین‌شده برای شرکت‌کننده نبوده، و در عوض شامل مواردی مانند انتظار داشتن احساس خوب باشد.

[491] descrimination
[492] Double-blind

محققی خلاق این امکان را بواسطه مقایسه‌ی میان اثرات رواندرمانی و اثرات یک سفر تفریحی به فلوریدا کشف کرد (McCardel, 1972).

در اینجا هدف برقراری تمایز درباره بخشهایی از افراد مورد هدف، موقعیتهای آزمایش، مقیاسهای اندازه گیری، مداخلهها، و زمانهایی است که در آنها رابطه علّت و معلولی موردنظر صدق نمی‌کند (برای مثال، اثر کاهش دادن اندازه کلاس درس هنگامی که کلاسهای جدید باید در مکانهای موقت مانند کانتینر برگزار شود، از میان می‌رود؛ یا اینکه اثر رواندرمانی تا شش ماه پس از پایان درمان قوی است، اما پس از پنج سال ناپدید می‌شود). اینگونه چالشها از این حقیقت ناشی می‌شوند که هر سازه می‌تواند چندین بخش داشته باشد. در اینجا، استراتژی آن است که سازه را به بخشهای کوچکتر سازنده آن که در بروز اثر مداخله اثرگذار هستند یا نیستند تقسیم نماییم. دوین و کوک (Devine & Cook, 1986) پاره‌ای از مطالعات را شناسایی کردند که میان نتایج مورد هدف (بهبود از جراحی) و دیگر سازه‌های کاملا مرتبط و مشابه تمایز برقرار نموده بودند. برای مثال، برخی مطالعات مدت زمانی را که پس از ترخیص از بیمارستان تا شروع فعالیتهای کاری و دیگر فعالیتهای عادی زندگی سپری شده بود را اندازه‌گیری کرده بودند، و این محاسبات نشان می‌داد که آموزش بیماران این دوران نقاهت را کاهش می‌دهد. هرچه اندازه نمونه بزرگتر باشد، احتمال یافتن متغیرهایی که به تفسیر خالص‌تر نتایج کمک می‌کنند، زیادتر می‌شود.

پاره‌ای اوقات، عکس حالت ذکر شده نیز رخ می‌دهد. به این معنی که بعضی واریانسها که جامعه‌ی محققین تا آن زمان آنها را مرتبط می‌دانستند، غیرمرتبط تشخیص داده می‌شوند. به عنوان مثال، جامعه محققین رواندرمانی بر این باور بودند که میزان تجربه درمانگر در نتایج مداخلات و مداخله اثرگذار است، و با افزایش تجربه درمانگر نتایج مشاهده شده در مراجعان نیز بهبود می یابد. با این حال، متاآنالیزهای انجام شده تا به حال، چنین اثری را گزارش نکرده‌اند (J. Berman & Norton, 1985; M. Smith et al., 1980; Shadish et al., 1993). چنین یافته‌هایی باعث می‌شود تا نقش چنین منابعی را مورد بازنگری قرار دهیم. برای مثال، با وجود اینکه تجربه درمانگر اثر اصلی و مستقیم تولید نمی‌کند اما آیا نمی‌تواند در تعامل با دیگر متغیرها (مانند شدت مشکل موجود) باشد؟ بدیهیست که یافته‌ها و باورهای پذیرفته‌شده فعلی در مورد عناصر مرتبط و یا غیرمرتبط همواره در طول زمان مورد بازنگری قرار می‌گیرند.

درون‌یابی[۴۹۳] و برون‌یابی[۴۹۴]

تعمیم دادن در حالتی با اطمینان و قوت بیشتر صورت می‌پذیرد که بتوان طیف افراد، شرایط آزمایش، مداخله‌ها، نتایج، و زمانهایی که در آن، رابطه علّت و معلولی یا به طور قوی و مستحکمی برقرار است، و یا اصلاً برقرار نیست را مشخص نمود. یافتن این شرایط نیازمند کاوش و کشف تجربی طیفهای موجود در مطالعات موجود است. در تحقیقات پزشکی، این مثالها انواع دوزها از صفر گرفته تا بالاترین دوز مجاز، طیف شرایط آزمون از بیمارستان ۵۰ تختخوابی گرفته تا ۵۰۰۰۰ تختخوابی، طیف بیماران سرطانی از سن ۲۵ سال گرفته تا سن ۷۵ سال، و طیف زمانی آزمونهای پیگیری از پس‌آزمون بلافاصله پس

[493] Interpolation
[494] Extrapolation

از درمان گرفته تا ۵ سال بعد از پایان درمان، را در بر می‌گیرد. در هر کدام از این موارد رتبه‌بندی انواع مثالهای سازه مقدور است و متاآنالیز در مقایسه با مطالعه تکی به نحوه بهتری این هدف تامین می‌نماید. در متاآنالیز، مقادیر افراطی و غیرمتعارف به فاصله‌های دورتری (با میزان کمتری) مجال بروز می‌یابند، و مقادیر معتدل و وسطی فراوانترند. در مقابل، در مطالعات تکی (بر مبنای یک مشاهده خاص) احتمال بیشتری وجود دارد که مقادیر بدست‌آمده در سرهای انتهایی و افراطی طیف قرار بگیرند.

به عنوان یک نمونه، در رابطه با میزان دوز مداخله، هوارد و همکارانش (Howard, Kopta, Krause, and Orlinsky, 1986) رابطه میان تعداد جلسات رواندرمانی و بهبود بیماران را بررسی نمودند. پانزده مطالعه جداولی را گزارش کردند که نشان می‌داد طول زمانهای بیشتر درمان، باعث ارتقاء وضعیت بیماران می‌شود. برای هر مطالعه، هوارد و همکارانش از تحلیل پروبیت برای نشان دادن درصد بهبود بیماران در جلسات صفر، ۱، ۲، ۴، ۸، ۱۳، ۲۶، ۵۲، و ۱۰۴ رواندرمانی (یک جلسه در هفته) استفاده کردند. یقیناً هیچکدام از مطالعه‌های تکی میزان بهبود در هر یک از این ۹ نقطه زمانی را گزارش نمی‌کرد. همچنین، هیچکدام از آنها درصد بهبود در جلسه صفر را گزارش نکرده بود، و هشت مطالعه پیگیری را پس از جلسه صدم متوقف کرده بودند. بنابراین، لازم بود تا تخمین بهبود برای این موارد [زمانی] برون‌یابی شود.

اگرچه برخی متاآنالیزهای صورت گرفته در مورد زمان پس‌آزمون نشان می‌دهند که اندازه اثر بدست آمده از رواندرمانی در پس‌آزمون تفاوت معناداری با اندازه‌های بدست آمده در مطالعات پیگیری ندارد، اما این نتیجه‌گیری تا حدی به برون‌یابی انجام شده از روی طولانی‌ترین پیگیری، به آنچه در صورت انجام مطاله پیگیری طولانی‌تر مشاهده می‌شد بستگی دارد. در رواندرمانی ازدواج و خانواده، متوسط زمان انجام پیگیری‌ها پنج ماه گزارش شده و طولانی‌ترین زمان پیگیری انجام شده تنها نه ماه بوده است. با وجود اینکه پیگیری‌هایی انجام شده با این طولهای زمانی نتایجی شبیه به اندازه اثر بدست آمده در پس‌آزمون بدست می‌دهند، شواهدی وجود دارد دال بر اینکه، اگر پیگیری‌هایی در یک سال بعد یا حتی مدتی دورتر صورت می‌گرفت، اندازه اثرات به میزان معنادارتری کاهش می‌یافت (Jocobson, Schmaling, & Holtzworth-Munroe, 1987; Snyder, Wills, & Grady-Fletcher, 1991). در نتیجه، برون‌یابی به مطالعات پیگیری با این طول زمانهای بسیار طولانی خطرساز است.

برون‌یابی و درون‌یابی غالباً با استفاده از برقراری همبستگی میان اندازه اثر و متغیر مورد نظر محقق بررسی می‌شود. مثلاً بکر (Becker, 1992) دریافت که همبستگی میان توانایی دانش‌آموزان و دستاوردهای علمی آنها مثبت است (برای زنان r = ۰/۳۳ و برای مردان r = ۰/۳۲) . به طور کلی می‌توان گفت که رابطه قوی مثبتی برقرار است. اما همبستگی‌های تجمیع شده خطی هستند، و برونیابی و درونیابی زمانی بهتر قابل انجام است که بتوان نموداری از همبستگیهای غیرخطی مشاهده کرد که در آن، امکان پیدا کردن انحرافها، شکستها و پیچشها در رابطه وجود داشته باشد. بکر در مطالعه خود قادر به انجام چنین کاری نبود، زیرا تمام مطالعات لحاظ شده در مطالعه وی همبستگی‌های خطی را گزارش کرده بودند. اما برخی اوقات در متاآنالیز محققین طیف‌هایی کمّی برای خودشان می‌سازند، و رابطه آن طیفها را با استفاده از تکنیکهای غیرخطی با اندازه اثر تعیین می‌کنند. بطور مثال، شدیش و همکارانش (Shadish et al., 2000) نشان دادند که اندازه اثرهای

مشاهده شده در مطالعات رواندرمانی تابعی از میزان نماینده‌بودن کلینیکی [۴۹۵] شرایطی بود که درمان در آن مورد بررسی قرار می‌گرفت. نماینده بودن کلینیکی برای هر مطالعه بر اساس ۱۰ گویه کدگذاری شد (برای مثال، آیا درمانگران حرفه‌ای مورد استفاده قرار گرفته بودند؟، آیا مراجعان از طریق منابع ارجاع استاندارد معرفی شده بودند؟). حاصلجمع نمرات بدست آمده از این گویه‌ها مجموع نمره نماینده‌بودن کلینیکی را (که در طیفی از ۱ تا ۱۰ بود) نشان می‌داد. نمودار پراکندگی رابطه میان اندازه اثر و نمرات بدست آمده از این گویه‌ها روندی خطی را نشان می‌داد؛ و افزودن عناصر غیرخطی به مدل آماری این داده ها اثر قابل‌توجهی در ارتقاء تخمین‌ها نداشت.

مطالعه شدیش و همکارنش(۲۰۰۰) همچنین نمایانگر کاربرد مفید یک روش آماری برای برونیابی و درونیابی در متاآنالیز است، که به آن مدلسازی سطح پاسخ [۴۹۶] گفته می‌شود (Rubin, 1990). در فصل قبل نحوه کاربرد این روش در مطالعات تکی مورد بحث قرار گرفت اما دامنه هر یک از متغیرها در یک مطالعه تکی محدودتر از آن است که بتوان این روش را برای آن به خوبی بکار گرفت. برای مثال بخش زیادی از مطالعات مورد بررسی در متاآنالیر شدیش تنها یک سطح از نمایندگی کلینیکی را بکار گرفته بودند. اما در متاآنالیز این دامنه به میزان قابل‌توجهی افزایش می‌یابد؛ در نتیجه، شدیش و همکارانش ۹۰ مطالعه را که از نظر سطح نمایندگی از بسیار اندک تا بسیار زیاد قرار داشتند را در مطالعه خود مورد بررسی قرار دادند. در این مطالعه، با استفاده از یک رگرسیون تصادفی، اثرات ابتدا اندازه اثر را از روی نمایندگی کلینیکی و مجموعه‌ای از متغیرهای همبسته و مزاحم مانند طراحی مطالعه، دوز درمان، مشخصات مقیاس اندازه‌گیری، و موقعیت گزارش از نظر انتشاراتی پیش‌بینی کردند. این تعدیلات از آن لحاظ ضروری بود که مطالعاتی که از نظر سطح نمایندگی کلینیکی بالاتر بودند احتمالاً به طور سیستماتیکی با مطالعات واجد سطح پایین‌تر متفاوت بودند. مثلاً، مطالعات با سطح بالاتر (در مقایسه با مطالعات با سطح پایین‌تر) کمتر احتمال داشت که از تخصیص تصادفی استفاده کرده باشند. پس از محاسبه اندازه اثر، محققین ضرایب رگرسیونی بدست آمده را برای برونیابی به نتایجی که در صورت انجام یک مطالعه ایده‌آل نمایندگی کلینیکی بدست می آمد، بکار گرفتند. منظور از مطالعه ایده‌آل در اینجا، مطالعه‌ای است که قویترین طرح آزمایشی را برای تخمین اثرات مداخله بکار گرفته، و بالاترین نمرات را از نظر نمایندگی کلینیکی بدست آورده باشد. هیچ مطالعه تکی‌ای وجود نداشت که در واقع واجد این شرایط باشد، اما مدل سطح پاسخ داده‌های در دسترس را برای (ساختن) نشان دادن نتایج آن مطالعه ایده آل مورد استفاده قرار داد. نتایج نشان داد که رواندرمانی در شرایط نماینده کلینیکی به همان میزان اثربخش است که در شرایط تحقیقاتی.

تبیین علّی

یافتن تبیین‌های علّی در متاآنالیز به سه روش انجام می‌شود. اول، در متاآنالیز محققین با سهولت بیشتری می‌توانند افراد، شرایط آزمون، مداخلهها، و نتایج را به اجزاء کوچکتر تجزیه نمایند؛ و از این طریق عناصر از نظر علّی مرتبط آنها را بیابند. به عنوان نمونه، دیوین و کوک (Devine & Cook, 1986) آموزش بیماران را به سه جزء تجزیه می‌کنند،که عبارت بود از ارایه اطلاعات، آموزش مهارتها، و پشتیبانی اجتماعی. اگرچه، هیچ مطالعه‌ای به تنهایی تمامی این ابعاد را در بر نمی‌گرفت. این

[495] Clinical representativeness
[496] Response surface modeling (RSM)

دو محقق، در مطالعه خود ترکیبهای متفاوت از این سه عنصر را ارزیابی کردند، و تحلیل های آنها نشان داد که استفاده از هر یک از عناصر به تنهایی، دارای کمترین اثربخشی است و ترکیب عناصر اندازه اثر را به صورت تجمیعی افرایش می‌دهد.
مهمترین محدودیت بکارگیری متاآنالیز برای یافتن عناصر مداخلات مرتبط علّی بیشتر کاربردی است تا نظری. این محدودیتها عبارتند از (۱) نبود یا کبمود اطلاعات و جزییات در خصوص عناصر مداخلات در بسیاری از نشریات و کتابها؛ (۲) مداخلات منتشر شده در مجلات بیشتر آن چیزی را نشان می‌دهند که محقق قصد انجام آن را داشته است، تا آن چیزی را که وی در حقیقت انجام داده است؛ (۳) نیاز به نمونه‌های با اندازه بزرگ، اگر لازم است تا تحلیلهای حساس به اندازه نمونه در مورد هر یک از اجزاء صورت گیرد. با این وجود، متاآنالیزهای فراوانی وجود دارد که به میزان قابل قبولی برای یافتن و تبیین اجزاء سازه‌های علّت و معلولی تلاش کرده‌اند.
دوم، در متاآنالیز می‌توان از رگرسیون چند–متغیره برای یافتن اطنابهای[۴۹۷] موجود میان متغیرهایی که نتایج مطالعه را تعدیل می‌کنند، سود برد. این کار به محدود کردن توضیحات و تبیین‌های علّی و ارزیابی بهتر اندازه و اهمیت اثر متغیرهای مستقل مختلف کمک می‌کند. برای مثال، لیپسی (Lipsey, 1992) نتایج صدها مطالعه انجام شده روی اثر مداخلههای مرتبط با جرایم مرتکب شده توسط نوجوانان را مورد بررسی قرار داد. وی بیش از صد متغیر مستقل اثرگذار بر ارتکاب این جرایم را کدگذاری کرد. نتایج تحلیل رگرسیون حامل یافته‌هایی بود، از قبیل اینکه تفاوت بیشتر میان گروهها در مرحله پیش‌آزمون با تفاوت بیشتر میان آنها در مرحله پس‌آزمون همبستگی داشت، اینکه مداخله‌های اعمال شده توسط محققین اثر بیشتری داشته، اینکه مداخله‌های اعمال شده در مکانهای عمومی مانند بازپروری‌ها نتایج ضعیفتری داشته؛ و یا اینکه مداخله‌های رفتاری و چندشکلی[۴۹۸] اثر قویتری داشته‌اند. برخی دیگر از متغیرهای پیش‌بین در تحلیل رگرسیون از نظر آماری غیرمعنادار بودند. از آن جمله می‌توان به غالب مشخصه‌های فردی نوجوانان و برخی از مداخله‌های عمومیت یافته مانند «ترس مستقیم» (که در آن به منظور دور کردن نوجوان از ارتکاب جرم وی را در معرض تصاویر ترسناکی از زندانیان که در محیطهای ناخوشایندی بودند قرار می‌دادند)، اشاره داشت.
سوم، تبیین و توضیح کامل نیازمند تحلیل فرایندهای ریز–میانجی‌گر علّی است، که پس از تغییر دادن علّت، و پیش از رخ دادن اثر حادث می‌شوند. مثال مرتبط با این موضوع را می‌توان در اثر انتظار روزنتال مشاهده کرد. روزنتال (Rosenthal, 1973) نظریه چهار عاملی اثر میانجی‌گری انتظار معلمین را ارائه داد. چهار عامل این نظریه عبارتند از (۱) شرایط جوی: انتظارات بالا در معلمین منجر به شکل‌گیری شرایط جوی اجتماعی–اقتصادی گرمتر می‌شود؛ (۲) بازخورد: انتظارات بالای معلمین منجر به بازخورد متفاوت‌تر در مورد درستی پاسخ دانش‌آموزان می‌شود؛ (۳) درونداد، معلمین مقادیر بیشتری مطلب و مطالب دشوارتر را به دانش‌آموزانی که نسبت به آنها انتظارات بالا دارند، درس می‌دهند؛ (۴) برونداد، معلمین به دانش‌آموزانی که از آنها انتظارات بالاتر دارند، شانس بیشتری برای پاسخ دادن می‌دهند. در سال ۱۹۸۵ هریس و روزنتال (Harris & Rosenthal, 1985) ۱۳۵ مطالعه را که در آنها رابطه میان انتظارات و این متغیرهای میانجی، و یا رابطه میان میانجی‌ها و نتایج مورد توجه قرار گرفته بود را بررسی کردند. تمامی اندازه اثرها به مقیاسهای مشترکی از همبستگی تبدیل

[497] Redundancy
[498] Multimodal

شده، و سپس برای هر یک از چهار عامل تجمیع شد. نتایج نشان داد که تمامی چهار عامل اثر موردنظر را میانجی‌گری کرده‌اند. با استفاده از این دانش بهتر می‌توانیم درک کنیم چطور اثر موردنظر را در موقعیتهای مختلف ایجاد کرده یا کاهش دهیم؛ و در نتیجه استنباطهای علّی تعمیم‌یافته را تقویت نماییم.

اگرچه، آزمون و بررسی فرایندهای میانجی‌گر علّی در متاآنالیز می‌تواند بسیار متفاوت باشد. یکی از مشکلات اصلی آن است که تعداد بسیار اندکی از مطالعات دست اول اثر میانجی‌ها را مورد بررسی قرار می‌دهند؛ و حتی زمانی که این کار را انجام می‌دهند نیز ندرتاً از مدلسازی‌های پیشرفته علّی یا روشهای متغیر ابزاری[۴۹۹] که در فصل پیش مورد بحث قرار گرفت استفاده می‌کنند. نتیجه آن است که با کمبود مطالعات مرتبط با میانجی‌گرها که از نظر تحلیل دقیق و به درستی انجام شده باشند روبرو هستیم (Becker, 1992). برخی مطالعات متاآنالیز، این مشکل را با درخواست از کدگذاران برای رتبه‌بندی مطالعات در مورد میانجیهای بالقوه دور می‌زنند (Shadish & Heinsman, 1997; Shadish & Ragsdale, 1996). اما این رتبه‌بندی‌های پس‌آزمونی و غیرمستقیم در مقایسه با زمانی که میانجی‌ها در مطالعات دست اول مورد محاسبه قرار می‌گیرند، از روایی و پایایی اندکی برخوردارند. گذشته از آن، برخی میانجی‌ها را نمی‌توان به این صورت رتبه‌بندی کرد. در نهایت، روشهای آماری قابل بکارگیری و سودمند برای مدلسازی میانجی‌ها برای متاآنالیز به خوبی تعدیل و مناسب‌سازی نشده‌اند. برای مثال، آماره های تطابق[۵۰۰] آنها غالباً نمی‌توانند اثرات تصادفی را در نظر بگیرند؛ و غالباً تنها با مدلهای اثرات ثابت تطابق می‌یابند. اگرچه کار بر روی این موضوعات در حال گسترش است و برنامه های کامپیوتری با قابلیت محاسبه اثرات تصادفی در حال طراحی و گسترش‌اند.

بحث در باب متاآنالیز

در سالهای ابتدایی معرفی متاآنالیز، این روش آماج انتقادات فراوانی قرار گرفت (Eysenck, 1978; Presby, 1978). برخی از این انتقادات سوالات پایه‌ای در مورد معناداری و ماهیت این روش مطرح کردند. یقیناً انجام‌دهندگان متاآنالیز خود در زمره دقیقترین و مصرترین منتقدین این روش قرار داشتند، و در تلاش بوده‌اند تا مشکلات اصلی همراه با این روش را مشخص کرده و راه‌حل‌هایی برای آنها بیابند. بنابراین، اشتیاق ما برای ارائه روش‌شناسی‌های جدید باید همرا با درک واقعگرایانه از محدودیتهای هر روش باشد.

علی‌رغم کاستی‌ها و ایرادات همراه با متاآنالیز، مشکلات همراه با مرور بر ادبیات‌ها غالبا جدی‌تر است. بنابراین، انتقادات وارده بر متاآنالیز باید با در نظر گرفتن این مسأله باشد که آیا مشکل موردنظر مختص متاآنالیز است، و یا در مورد هر مرور بر ادبیات دیگری نیز صادق است. و اینکه آیا مشکل موردنظر به اندازه‌ای جدی است که ما را مجاب به بازگشت به تمامی مشکلات ناشی از تکنیکهای مرور بر ادبیات کند؟ برای مثال، کریمر و همکارانش (Kraemer et al., 1998) پیشنهاد می‌کنند که مطالعات با توان آماری پایین از متاآنالیز حذف شوند، زیرا این کار می‌تواند سوگیری ناشی از مشکل کشوی میزکار و همچنین سوگیری ناشی از نتایجی که تنها به عنوان غیرمعنادار گزارش شده‌اند را کاهش دهد. اگرچه در برخی

[499] Instrumental variable methods
[500] Fit statistics

حوزه‌های ادبیات، یافتن مطالعاتی که از توان آماری کافی برخوردارند (بیش از ۰/۸) دشوار است. با این حال در همین حوزه‌ها نیز همچنان مرور بر ادبیات انجام می‌شود؛ و این مرور بر ادبیات‌های متنی تمامی ایراداتی که متاآنالیز در تلاش برای رفع آن است را با خود دارند. در نتیجه، راه‌حل‌های جایگزینی که توصیه به حذف ادبیات با توان آماری پایین می‌کنند، و یا تنها بر مرور بر ادبیات‌های متنی تکیه می‌کنند، نه تنها غیر واقعگرایانه است، بلکه درمانیست بدتر از مشکل.

با گذشت زمان نقدهای جدید در خصوص متاآنالیز مطرح می‌شود. اگرچه برخی از آنها در برهه‌ای از زمان توجه محققین را به خود جلب می‌کنند، بعضاً واجد روایی چندانی نیستند. برای مثال، یافته‌های للوریه و همکارانش (Lelorier et al., 1997) تمایز و تفاوت جدی میان نتایج بدست‌آمده از متاآنالیز آزمایشهای کوچکتر، و آزمایشهای تکی تصادفی بزرگ نشان داد. اگرچه، مشکلات مورد بحث در مطالعه آنها از بسیاری جهات قانع‌کننده نبود. برای مثال، بکارگیری مدلهای اثر ثابت به جای مدلهای اثر تصادفی باعث افزایش خطای نوع یک در مطالعه آنها شده بود، و یا اینکه روش‌های بکار گرفته شده در مطالعه آنها، از برخی جهات به شکل ضعیفی تبیین شده بود (مثلاً اینکه آیا متاآنالیز انجام شده آزمایش‌های غیرتصادفی را نیز در بر گرفته بود؟). در همین راستا، مطالعه دیگری که مشکلات فوق‌الذکر را برطرف کرده و مطالعه للوریه و همکارانش را تکرار کرده نتایج کاملاً متضاد با آنها بدست آورد. به این معنی که این مطالعه نشان داد که متاآنالیز آزمایشهای کوچک، نتایجی شبیه آزمایشهای بزرگ تصادفی به دست می‌دهد.

در نهایت، توصیه نگارندگان به خوانندگان این کتاب آن است که با همان نگاه انتقادی‌ای به متاآنالیز نگاه کنید که به دیگر روشهای آماری می‌نگرید. متاآنالیز روش بسیاری جدیدی است، بنابراین هر روز انتقادهایی نسبت به آن مطرح می‌شود و در پاسخ راه‌حل‌هایی ارائه می‌شود. این حالت برای هر روش جدیدی در علم پیش می‌آید. اما این موضوع نباید باعث نادیده گرفتن این واقعیت شود که اضافه کردن روشهای کمّی به مجموعه تکنیکهای انجام مرور بر ادبیات، یکی از مهمترین رشدهای رخ داده در علوم اجتماعی در نیمه دوم قرن بیستم بوده است.

References

Abadzi, H. (1984). Ability grouping effects on acildemic achievement and self-esteem in a southwestern school district. Journal of Educational Research~ 77, 287-292.

Abad~i, H· (1985). Ability grouping effects on i)Cademic achievement and self-esteem: Who performs in the long run as expected? Journal of Educational Research, 79, 36-39.

Abelsol), R. P. (1995). Statistics as principled argument. Hillsdale, NJ: Erlbaum. . Abelson, R. P. (1996). Vulnerability of contrast tests to simpler interpretations: An addendum to Rosnow and Rosenthal. Psychological Science, 7, 242-246.

Abelson, R. P. (1997). On the surprising longevity of flogged ho.rses: Why there is a case for the significance test. Psychological ~cience, 8, 12-15.

Achen, C. H. (1986). The statistical analy;is of quasi-experiments. Berkeley: University of California Press. Ad;lir, J. G. (1973). The Hawthorne effect: A reconsideration of a methodological artifact. Journal of Applied Psychology, 69, 334-345.

Ahlborn, A., & Norell, S. (1990). Introduction to modern epidemiology. Chestnut Hill, MA: Epidemiology Resources. Ahn, C.-J(. (1983). A Monte Carlo comparison of statistical methods for estimating treatment eff~ct~ in regression discontinuity design (Doctoral dissertation, Washington State University, 1983).·Dissertation Abstracts International, 44(03), 733A.

Aigner, D.]., & Hausman, J. A. (1980). Correcting for truncation bias in the analysis of experiments in time-of-day pricing of electricity. Bell Journal, 35, 405. Aiken, L. S., & West, S. G. (1990). Invalidity of true experiments: Self-report pretest biases. Evaluation Review, 14, 374-390.

Aiken, L. S., & West, S. G. (1991). Testing and interpreting interactions in multiple regression. Newbury Park, CA: Sage.

Aiken, L. S., West, S. q., ~chwahn, D. E., Carroll, J. L., & Hsiung, S. (1998). Comparison of a randomized and two quasi-experimental designs in a single outcome evaluation: Efficacy of a university-level remedial writing program. Eyaluation Review, 22, 207-244. Albermarle Paper Co. v. Moody, 442 U.S. 435 (1975).

Alexander, R. A., Barrett, G. V., Alliger, G. M., & Carson, J(. P. (1986). Towards a general model of uon-random sampling and the impact on population correlations: Generalization of Berkson's Fallacy and restriction of range. British Journal of Mathematical and Statistical Psychology, 39, 90-115.

Allen,]. P., Philliber, S., Herr ling, S., &'Kuperminc, G. P. (1997). Preventing teen pregnancy and academic failure: Experimental evaluation of a developmentally based approach. Child Development, 64, 729-742.

Allison, D. B. (1995). When is it worth measuring a covariate in a randomized trial? Journal of Consulting and Clinical Psychology, 63, 339-343.

Allison, D. B., Allison, R. L., Faith, M.S., Paultre, F., & Pi-Sunyer, F. X. (1997). Power and money: Designing statistically powerful studies while minimizing financial costs. Psychological Methods, 2, 20-33.

Allison, D. B., & Gorman, B.S. (1993). Calculating effect sizes for meta-analysis: The case of the single case. Behavior Research and Therapy~ 31, 621-631.

Allison, P. D. (1987). Estimation of linear models with incomplete data. In C. Clogg (Ed.), Sociological methodology (pp. 71-103). San Francisco: Jossey-Bass.

Allison, P. D., & Hauser, R. M. (1991). Reducing bias in estimates of linear models by remeasurement of a random subsample. Sociological Methods and Research, 19, 466-492.

Alwin, D. F., & Tessler, R. C. (1985). Causal models, unobserved variables, and experimental data. In H. M. Blalock (Ed.), Causal models in panel and experimental designs (pp. 55-88). New York: Aldine.

Amato, P.R., & Keith, B. {1991). Parental divorce and the well-being of children: A metaanalysis. Psychological Bulletin, 110, 26-46.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. {1985). Standards for educational and psychological testing. Washington, DC: American Psychological Association.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. {1999). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

American Evaluation Association. (1995). Guiding principles for evaluators. In W. R. Shadish, D. L. Newman, M.A. Scheirer, & C. Wye (Eds.), Guiding principles for evalu· ators (pp. 19-26). San Francisco: Jossey-Bass.

American Psychiatric Association. {1994). Diagnostic and statistical manual of mental disorders (4th ed.). Washington, DC: Author. American Psychiatric Association. (2000). Handbook of psychiatric measures. Washington, DC: Author.

Arllerican Psychological Association (1954). Technical recommendations for psychological tests and diagnostic techniques. Psychological Bulletin, 51 {Supplement).

American Psychological Association. (1992). Ethical' principles of psychologists and code of conduct. American Psychologist, 47, 1597-1611.

American Psychological Association. (1994). Publication manual of the American Psychological Association (4th ed.). Washington DC: Author.

Anastasi, A. (1968). Psychological testing (3rd ed.). New York: Macmillan. Anderman, C., Cheadle, A., Curry, S., Diehr, P., Shultz, L., & Wagner, E. (1995). Selection bias related to parentalconsentinschool-basedsurvey research. Evaluation Review, 19, 663-674.

Anderson, C. A., Lindsay, J.]., & Bushman, B. J. (1999). Research in the psychological lab· oratory: Truth or triviality? Current Directions in Psychological Science, 8, 3-10.

Anderson,]. G. {1987). Structural equation models in the social and behavioral sciences: Model building. Child Development, 58, 49-64.

Anderson, V. L., & McLean, R. A. (1984). Applied factorial and fractional designs. New York: Dekker.

Angrist,]. D., & Imbens, G. W. (1995). ~wa-stage least squares estimation of average causal effects in models with variable treatment intensity. Journal of the American Statistical Association, 90, 431-442.

Angrist, J.D., Imbens, G. W., & Rubin, D. B. (1996a). Identification of causal effects using instrumental variables.] ournal of the American Statistical Association, 91, 444--4 55.

Angrist, J.D., Imbens, G. W., & Rubin, D. B. (1996b). Rejoinder. Journal of the American Statistical Association, 91, 468-472.

Anson, A., Cook, T. D., Habib, F., Grady, M. K., Haynes, N. & Comer, J.P. (1991). The Comer School Development Program: A theoretical analysis. Journal of Urban Education, 26, 56-82.

Arbuckle, J. J. {1997). Amos users' guide, version 3.6. Chicago: Small Waters Corporation. Armitage, P. (1999). Data and safety monitoring in the Concorde and Alpha trials. Controlled Clinical Trials, 20, 207-228.

Aronson, D. (1998). Using sibling data to estimate the impact of neighborhoods on children's educational outcomes. Journal of Human Resources~ 33, 915-956.

Ashenfelter, 0. (1978). Estimating the effects of trainiug programs on earnings. Review of Economics and Statistics, 60, 47-57.

Ashenfelter, 0., & Card, D,. (1985). Using the longitudinal structure of earnings to estimate the effect of trainiug programs. Review of Economics and Statistics~ 67, 648-660.

Ashenfelter, 0., & Krueger, A. B. (1994). Estimates of the economic returns to schooling from a new sample of twins. American Economic Review, 84, 1157-1173.

Atkinson, A. C. (1985). An introduction to the optimum design of experiments. In A. C. Atkinson & S. E. Fienberg (Eds.), A celebration of statistics: The lSI centenary volume (pp. 465-473). New York: Springer-Verlag.

Atkinson, A. C., & Donev, A. N. (1992). Optimum experimental designs. Oxford, England: Clarendon Press.

Atkinson, D. R., Furlong, M. J., & Wampold, B. E. (1982). Statistical significance, reviewer evaluations, and the scientific process: Is there a (statistically) significant relationship? Journal of Counseling Psychology, 29, 189-194.

Atkinson, R. C. (1968). Computerized instruction and the learning process. American Psychologist, 23, 225-239.

Atwood, J, R., & Taylor, W. (1991). Regression discontinuity design: Alternative for nursing research. Nursing Research, 40, 312-315.

Babcock, J. L. (1998). Retrospective pre~ests: Conceptual and methodological issues (Doctoral dissertation, University of Arizona, 1997). Dissertation Abstracts International, 58(08), 4513B.

Bagozzi, R. P., & Warshaw, P.R. (1992). An examination of the etiology of the attitudebehavior relation for goal-directed behaviors. Multivariate Behavioral Research, 27, 601-634.

Baker, F., & Curbow, B. (1991). The case-control study in health program evaluation. Evaluation and Program Planning, 14,263-272.

Baker, S. H., & Rodriguez, 0. (1979). Random time quote selection: An alternative to random selection in experimental evaluation. In L. Sechrest, S. G. West, M. A. PhillipRednet; & W. Yeaton (Eds.), Evaluation studies review annual (Vol. 4, pp. 185-196).

Beverly Hills, C< Encino, CA: Dickenson. Bechtel, W. (1988). Philosophy of science: An overview for cognitive science. Hillsdale, K} Erlbaum.

Beck, A. T., Ward, C. H., Mendelsohn, M., Mock,]., & Erbaugh, J. (1961). Anlnventory for measuring depression. Archi~·es of General Psychuttry, 4, 561-5'71.

Becke~ B. J, 11988). Synthesizing standardized mean-change measures. British Journal of Mathematical and Statistical Psychology, 41, 257-278.

Becker, B. J. (1992). Models of sdence achievement Forces aff~iing male and female performance in school science. InT. D.

Cook, H. M. Cooper, D. S. Cordray,_ H. Hartmann, L. V. Hedges, R. J. Ligbt, T. A. Louis, & F. Mosteller (Eds.), Meta-analysis for explana· tion: A casebook (pp. 209-281). New York: Russell Sage Foundation.

Bocke~ B. J. (1994). Combining significance levels. In H. Cooper & L. V. Hedges (Eds.), The handbook of research synthesis (pp. 215-230). New York: Russell Sage Foundation.

Becker, B. J., & Schram, C. M. (1994), Examining explanatory models through research synthesis. Jn H. Cooper & L. V, Hedges (Eds.), The handbook of research synthesis (pp. 357-381). New York: Russell Sage Foundation.

Becker, H. S. (1958}. Problems of Inference and proof in participant observation. American Sociological Review, 23, 652-<'60.

Becker, H. S. (1979). Do photographs tell the truth? InT. D. Cook & C. S. Reichardt (Eds.), Qualitative and quantitative methods in evaluation research (pp. 99-117). London: Sage.

Becker, M. H. (1992). Theoretical models of adherence and strategies for improving adherence. InS. A. Shumaker, E. B. Schron, & J. K. Onkene (Eds.), The handbook of health behavior change (pp. 5-43). New York: Springer.

Beecher, H. (1955). The powerful placebo. Journal of the American Medical Association, 159, 1602-1606.

Beecher, H. (1966). Ethics and clinical research. New England Journal of Medicine, 274, 1354-1360.

Begg, C., Cho, M., Eastwood, S., Horton, R., Moher, D., Olkin, I., Pitkin, R., Rennie, D., Schulz, K. F., Simel, D., & Stroup, D. F. (1996). Improving the quality of reporting of randomized controlled trials. Journal of the American Medical Association~ 276, 637-639.

Begg, C. B. (1990). Suspended judgment: Significance tests of covariate imbalance in clinical trials. Controlled Clinical Trials, 11,223-225.

Begg, C. B. (1994). Publication bias. In H. Cooper & L. V. Hedges (Eds.), The handbook of research synthesis (pp. 399-409). New York: Russell Sage Foundation.

Begg, C. B. (2000). Ruminations on the intent-to-treat principle. Controlled Clinical Trials, 21, 241-243.

Bell, S. H., Orr, L. L., Blomquist,}. D., & Cain, G. G. (1995). Program applicants as a comparison group in evaluating training programs. Kalamazoo, Ml: Upjohn Institute for Employment Research.

Bentler, P.M. (1987). Drug use and personality in adolescence and young adulthood: Structuralmodels with nonnormal variables. Child Development, 58, 65-79.

Bentler, P.M. (1993). EQS!Windows user's guide. (Available from BMDP Statistical Software, Inc., 1440 Sepulveda Blvd., Suite 316, Los Angeles, CA 90025)

Bentle~ P.M. (1995). EQS: Structural equations program manual. Encino, CA: Multivariate Software.

Bentler, P. M,. & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. Psychological Bulletin, 88~ 588-606.

Bentler, P.M., & Chou, C.-P. (1988). Practical issues in structural modeling. In J. S. Long (Ed.), Common problems/proper solutions: Avoiding error in quantitative research (pp. 161-192). Newbury Park, Ck Sage.

Bentler, P.M., & Chou, C.-P. (1990). Model search with TETRAD II and EQS. Sociological Methods and Research, 19, 67-79.

Bentler, P.M., & Speckart, G. (1981). Attitudes "cause" behaviors: A structural equation analysis. Journal of Personality and Social Psychology, 40, 226-238. Bentler, P.M., & Woodward, J. A. (1978). A Head Start reevaluation: Positive effects are not yet demonstrable. Evaluation Quarterly, 2, 493-510.

Bentler, P.M., & Wu, E. J. C. (1995). EQS/Windows user's guide. (Available from Multivariate Software, Inc., 4924 Balboa Blvd., #368, Encino, CA 91316) Berg, A. T., & Vickrey, B. G. (1994). Outcomes research. Science, 264, 757-758.

Berg, B. L. (1989). Qualitative research methods for the social sciences. Boston: Allyn & Bacon.

Berger, V. W., & Exner, D. V. (1999). Detecting selection bias in randomized clinical trials. Controlled Clinical Trials, 20, 319-327.

Bergner, R. M. (1974). The development and evaluation of a training videotape for the resolution of marital conflict. Dissertation Abstracts International, 34, 3485B. (University Microfilms No. 73-32510).

Berk, R. A., & DeLeeuw, J. {1999). An evaluation of California's inmate classification system using a generalized regression discontinuity design. Journal of the American Statis~ tical Association, 94, 1045-105~.

Berk, R. A., Lenihan, K. J., & Rossi, P. H. (1980). Crimf and poverty: Some experimental evidence from ex~offenders. American Sociological Review, 45, 766-786.

Berk, R. A., & R;mma, D. (1983). Capitalizing on nonrandom assignment to treatment: A regression discontinuity evaluation of a crime control program. Journal of the American Statistical Association, 78, 21-27.

Berk, R. A., Smyth, G. K., & Sherman, L. W. (1988). When random assignment fails: Some lessons from the Minneapolis Spouse Abuse Experiment. Journal of Quantitative Criminology, 4, 209-223.

Berman, J. S., & Norton, N.C. (1985). Does professional training make a therapist more effective? Psychological Bulletin, 98, 401-407.

Berman, P., & McLaughlin, M. W. (1977). Federal programs supporting educational change: Vol. 8. Factors affecting implementation and continua~jon. Santa Monica, CA: RAND.

Besadur, M., Graen, G. B., & Scandura, T. A. (1986). Training effects on attitudes toward divergent thinking among manufacturing engineers. Journal of Applied Psychology, 71, 612-617.

Beutler, L. E., & Crago, M. (Eds.). (1991). Psychotherapy research, An international review of programmatic studies. Washington, DC: American Psychological Association.

Bhaskar, R. (1975). A realist theory of science. Leeds, England, Leeds. Bickman, L. (1985). Randomized field experiments in ec!-ucation: Implementation lessons. In R. F. Boruch & W. Wothke (Eds.), Randomization and field experimentation (pp. 39-53). San Francisco' Jossey-Bass.

Biglan, A., Hood, D., Brozovsky, P., Ochs, L., Ary, D., & Black, C. (1991). Subject attri~ tion in prevention research. In C. G. Leukfeld & W. Bukowski (Eds.), Drug use prevention interoention research: Methodological issues (NIDA Research Monograph 107, DHHS Publication No. 91-1761, pp. 213-228). Rockville, MD' U.S. Government Printing Office.

Biglan, A., Metzler, C. W., & ALy, D. V. (1994). Increasing the prevalence of successful children: The case for community intervention research. Behavior Analyst, 17, 335-351.

Birnbaum, A. (1961). Confidence curves: An omnibus technique for estimation and testing statistical hypotheses. Journal of the American Statistical Association, 56, 246-249.

Bishop, R. C., & Hill, J. W. (1971). Effects of job enlargement and job change on contiguous but non-manipulated jobs as a function of worker's status. Journal of Applied Psychology, 55, 175-181.

Blackburn, H., Luepker, R., Kline, F. G., Bracht, N., Car law, R., Jacos, D., Mittelmark, M., Stauffer, L., & Taylor, H. L. {1984). The Minnesoq Heart Health Program: A research and demonstration project in cardiovascular disease prevention. In J. D. Matarazzo, S. Weiss,]. A. Herd, N. E. Miller, & S.M. Weiss (Eds.), Behavioral health (pp. 1171-1178). New York: Wiley.

Bloom, B. S. (Ed.). (1956). Taxonomy of educational objectives: Handbook I. The cognitive ~amain. New York: ~cKay.

Bloolll~ H. S. (1984a). Accounting for no-shows in experimental evaluation designs. Evaluatiqn Review~ 8~ 225-246.

Bloom; :H. S. (1984b). Estimating the effect of job-training programs, using longitudinal data: Ashenfelter's findings reconsidered. Journal of Human Resources, 19, 544-556.

Bloo.p1, H. S. (1990). Back to work: Testing reemployment services for displaced workers. Kalamazoo, MI: Upjohn Institute.

Bloom, H. S., & Ladd, H. F. (1982). Property tax revaluation and tax levy growth. Journal of Urban Economics, 11, 73-84.

Bloor, D. (1976). Knowledge and social imagery. London: Routledge & Kegan Paul. Bloor, D. (1997). Remember the strong prograrii? Science, Technology, and Human Values, 22, 373-385.

Bock, R. D. (Ed.). (1989). Multilevel analysis of educational data. San Diego, CA: Academic Press.

Boissel,J. P., Blanchard,]., Panak, E., Peyrieux,J. C., & Sacks, H. (1989). Considerations for the meta-analysis of randomized clinical trials. Controlled Clinical Trials, 10, 254-281.

Boll~n, K. A. (1989). Structural equations with latent variables. New York: Wiley.

Bollen, K. A. (1990). Outlier screening and a distribution-free test for vanishing tetrads. Sociological Methods and Research, 19, 80-9,2.

Bo0msma, A. (1987). The-robustness of maximum likelihood estimation in structural equation models. In P. Cutrance & R. Ecob (Eds.), Structural modeling by example: Applications in educational, sociological, and behavioral research (pp. 160-188). Cambridge, England: Cambridge University Press.

Borenstein, M., & Coheh,J. (1988). Statistical power analysis: A computer program. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Borenstein, M., Cohen,J., & Rothstein, H. (in press). Confidence intervals, effect size, and power [Computer program]. Hillsdale, NJ: Erlbaum.

Borenstein, M., & .Rothstein, H. (199.9). Comprehensive meta-analysis. Englewood, NJ: Biostat.

Borkovec, T. D., & Nau, S.D. (1972). Credibility of analogne therapy rationales. Journal of Behavior Th;,.apy and Experimental Psychiatry, 3, 257-260.

Boruch, R. F. (1975). Coupling ra~domized experiments and approximations ro experiments in social program evaluation, Sociological Methods and Research, 4, 31-53.

Boruch, R. F. (1982). Experimental tests in education: R~~ommendations from the Holtzman Report. American Statistician, 36, 1-8.

Boruch, R. F. (1997). Randomized field experiments for planning and evaluation: A practical guide. Thousand Oaks, CA: Sage.

Boruch, R. F., & Cecil, J. S. (1979). Assuring the confidentiality of social research data. Philadelphia: University of Pennsylvania Press.

Boruch, R. F., Dennis, M., & Carter-Greer, K. (1988). Lessons from the Rockefeller Foundation's experilnents on the Minority Female Single Parent program. Evaluation Review, 12, 396-426.

Boruch, R. F., & Foley, E. (2000). The honestly experimental society: Sites and other entities as the units of allocation and analysis in randomized trials. In L. Bickman (Ed.), Validity and social experimentation: Donald Campbell's legacy (Vol. 1, pp. 193-238). Thousand Oaks, CA: Sage.

Boruch, R. F., & Gomez, H. (1977). Sensitivity, bias, and theory in impact evaluations. Professional Psychology, 8, 411-434.

Boruch, R. F., &Wothke, W. (1985). Seven kinds of randomization plans for designing field experiments. In R. F. Boruch & W. Wothke (Eds.), Randomization and field experimentation (pp. 95-118). San Francisco: Jossey-Bass.

Bos, H., Huston, A., Granger, R., Duncan, G., Brock, T., & McLoyd, V. (1999, April). New hope for people with low incomes: Two-year results of a program to reduce poverty and reform welfare. New York: lvianpower Research Development.

Box, G. E. P., & Draper, N. R. (1987). Empirical model-building and response surfaces. New York: Wiley.

Box, G. E. P., Hunter, W. G., & Hunter, J. S. (1978). Statistics for experimenters: An introduction to design, data analysis, and model building. New York: Wiley.

Box, G. E. P., & Jenkins, G. M. (1970). Time series analysis: Forecasting and control. San Francisco: Holden-Day.

Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (1994). Time series analysis: Forecasting and control (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Bracht, G. H., & Glass, G. V. (1968). The external validity of experiments. American Educational Research journal, 5, 437-474.

Bra"dburn, N. M. (1983). Response effects. In P. H. Rossi, J.D. Wright, & A. B. Anderson (Eds.), Handbook ofsurvey research (pp. 289-328). San Diego, CA: Academic Press.

Braden, J. P., & Bryant, T. J. (1990). Regression discontinuity designs: Applications for school psychologists. School Psychology Review, 19, 232-239.

Bramel, D., & Friend, R. (1981). Hawthorne, the myth of the docile worker, and class bias in psychology. American Psychologist, 36, 867-878.

Braught, G. N., & Reichardt, C. S. (1993). A computerized approach to trickle-process, random assignment. Evaluation Review, 17, 79-90.

Braunholtz, D. A. (1999). A note on Zelen randomization: Attitudes of parents participating in a neonatal clinical trial. Controlled Clinical Trials, 20, 569-571.

Braver, M. C. W., & Braver, S. L. (1988). Statistical treatment of the Solomon Four-Group design: A meta-analytic approach. Psychological Bulletin, 104, 150-154.

Braver, S. L., & Smith, M. C. (1996). Maximizing both external and internal validity in longitudinal true experiments with voluntary treatments: The "combined modified" design. Evaluation and Program Planning, 19, 287-300.

Breger, M.J. (1983). Randomized social experiments and the law. In R. F. Boruch & J. S. Cecil (Eds.), Solutions to ethiwl and legal problems in social research (pp. 97-144). New York: Academic Press.

Breslau, D. (1997). Contract shop epistemology: Credibility and problem construction in applied social science. Social Studies ofScience, 27, 363-394.

Brockwell, P. J., & Davis, R. A. (1991). Time series: Theory and methods (2nd ed.). New York: Springer-Verlag.

Bronfenbrenner, U. (1979). The ecology of human development: Experiments by nature and design. Cambridge, Massachusetts: Harvard University Press.

Brook, J. S., Cohen, P., & Gordon, A. S. (1983). LongitUdinal study of adolescent drug use. Psychological Reports, 53, 375-378.

Brown, R. (1986). Social psychology: The second.edition. New York: Free Press.

Brown, H. I. (1977). Perception, theory and commitment: The new philosophy of science.

Chicago: University of Chicago Press.

Brown, H. I. (1989). Toward a cognitive psychology of What? Social Epistemology, 3, 129-138.

Brunette, D. (1995). Natural disasters and commercial real estate returns. Real Estate Finance, 11, 67-72.

Brunswik, E. (1956). Perception and the representative design of psychological experiments (2nd ed.). Berkeley: University of California Press.

Bryk, A. S., & Raudenbush, S. W. (1992). Hierarchical linear models: Appliwtions and data analysis methods. Newbury Park, CA: Sage.

Bryk, A. S., Raudenbush, S. W., & Congdon, R. T. (1996). HLM: Hierarchical linear modeling with the HLM/2L and HLM/3L programs. (Available from Scientific Software International, 1525 E. 53rd Street, Suite 530, Chicago IL 60615)

Bunge, M. (1959). Causality and modern science (3rd ed.). New York: Dover.

Bunge, M. (1992). A critical examination of the new sociology of science (Part 2). Philosophy of the Social Sciences, 22, 46-76.

Burger, T. (in press). Ideal type: Understandings in the social sciences. InN. Smelser & P. Baltes (Eds.), Encyclopedia of the behavioral and social sciences. Amsterdam: Elsevier.

Burtless, G. (1995). The case for randomized field trials in economic and policy research. Journal of Economic Perspectives, 9, 63-84. .

Byrne, B. (1989). A primer ofLISREL. New York: Springer-Verlag.

Byrne, B. (1994). Structural equation modeling with EQS and EQS/Windows: Basic concepts, applications, and programming. Newbury Park, CA: Sage.

Cahan, S., & Davis, D. (1987). A between-grade-levels approach to the investigation of the absolute effects of schooling on achievement. American Educational Research journal, 24, 1-12.

Cahan, S., Linchevski, L., Ygra, N., & panziger, I. (1996). The cumulative effect of ability grouping on mathematical achievement: A longitudinal perspective. Studies in Educational Evaluation, 22, 29-40.

Cain, G. G. (1975). Regression and selection models to improve nonexperimental comparisons. In C. A. Bennett & A. A. Lumsdaine (Eds). Evaluation and experiment: Some critical issues in assessing social programs (pp. 297-317). New York: Academic Press.

Caines, P. E. (1988). Linear stochastic systems. New York: Wiley.

Campbell, D. T. (1956). Leadership and its effects on groups (Ohio Studies in Personnel, Bureau of Business Research Monograph No. 83). Columbus: Ohio State University. Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. Psychological Bulletin, 54, 297-312.

Campbell, D. T. (1963). From description to experimentation: Interpreting trends as quasi-experiments. In C. W. Harris (Ed.), Problems in measuring change (pp. 212-243). Madison: University of Wisconsin Press.

Campbell, D. T. (1966a).Pattern matching as an essential in distal knowing. InK. R. Hammond (Ed.), The psychology of Egon Brunswik. New York Holt, Rinehart, & Winston.

Campbell, D. T. (1966b). The principle ofproximal similarity in the application ofscience. Unpublished manuscript, Northwestern University.

Campbell, D. T. (1969a). Prospective: Artifact and control. In R. Rosenthal & R. L. Rosnow (Eds.), Artifact in behavioral research (pp. 351-382). New York Academic Press.

Campbell, D. T. (1975). "Degrees of freedom" and the case study. Comparative Political Studies, 8, 178-193.

Campbell, D. T. (1976). Focal local indicators for social program evaluation. Social Indicators Research, 3, 237-256.

Campbell, D. T. (1978). Qualitative knowing in action research. In M. Brenner & P. Marsh (Eds.), The social contexts ofmethod (pp. 184--209). London: Croom Helm.

Campbell, D. T. (1982). Experiments as arguments. In E. R. House (Ed.), Evaluation studies review annual (Volume 7, pp. 117-127). Newbury Park, CA: Sage.

Campbell, D. T. (1984). Foreword. In W. M. K. Trochim, Research design for program evaluation: The regression discontinuity approach (pp. 15-43). Beverly Hills, CA: Sage.

Campbell, D. T. (1986). Relabeling internal and external validity for applied social scientists. In W. M. K. Trochim (Ed.), Advances in quasi-experimental design and analysis (pp. 67-77). San Francisco: Jossey-Bass.

Campbell, D. T. (1988). Methodology and eptiStemology for social science: Selected papers (E. S. Overman, Ed.). Chicago: University of Chicago Press.

Cimpbell, D. T., & Boruch, R. F. (1975). Making the case for randomized assigument to treatments by considering the alternatives: Six ways in which quasi-experimental evaluations in compensatory education tend to underestimate effects. In C. A. Bennett & A. A. Lumsdaine (Eds.), Evaluation and experiments: Some critical issues in assessing social programs (pp. 195-296). New York: Academic Press.

Campbell, D. T., & Erlebacher, A. E. (1970). How regression artifacts can mistakenly make compensatory education programs look harmful. In j. Hellmuth (Ed.), The Disadvantaged Child: Vol. 3, Compensatory education: A national debate (pp. 185-210). New York: Brunner/Mazel.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 56, 81-105.

Campbell, D. T., & Kenny, D. A. (1999). A primer on regression artifacts. New York: Guilford Press.

Campbell, D. T., & Russo, M. J. (1999). Social experimentation. Thousand Oaks, CA: Sage.

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research. Chicago: RandMcNally.

Campbell, F. A., & Ramey, C. T. (1995). Cognitive and school outcomes for higb-risk African American students at middle adolescence: Positive effects of early intervention. American Educational Research journal, 32, 743-772.

Camstra, A., & Boomsma, A. (1992). Cross-validation in regression and covariance structure analysis. Sociological Methods and Research, 21, 89-115.

Canner, P. (1984). How much data should be collected in a clinical trial? Statistics in Medicine, 3, 423-432.

Canner, P. (1991). Covariate adjustment of treatment effects in clinical trials. Controlled Clinical Trials, 12, 359-366.

Capaldi, D., & Patterson, G. R. (1987). An approach to the problem of recruitment and retention rates for longitudinal research. Behavioral Assessment, 9, 169-177.

Cappelleri, J. C. (1991). Cutoff-based designs in comparison and combination with randomized clinical trials. UnpUblished doctoral dissertation, Cornell University, Ithaca, New York.

Cappelleri, J. C., Darlington, R. B., & Trochim, W. M. K. (1994). Power analysis of cutoff-based randomized clinical trials. Evaluation Review, 18, 141-152.

Cappelleri, J. C., Ioannidis, J.P. A., Schmid, C. H., deFerranti, S.D., Aubert, M., Chalmers, T. C., & Lau, J. (1996). Large trials vs meta-analysis of smaller trials: How do their results compare? Journal of the American Medical Association, 276, 1332-1338.

Cappelleri, J. C., & Trochim, W. M. K. (1992, May). An illustrative statistical analysis of cutoff-based randomized clinical trials. Paper presented at the annual meeting of the Society for Clinical Trials, Philadelphia, PA.

Cappelleri, J. C., & Trochim, W. M. K. (1994). An illustrative statistical analysis of cutoff-based randomized clinical trials. Journal of Clinical Epidemiology, 47, 261-270.

Cappelleri, J. C., & Trochim, W. M. K. (1995). Ethical and scientific features of cutoff- based designs of clinical trials. Medical Decision Making, 15, 387-394.

Cappelleri, J. C., Trochim, W. M. K., Stanley, T. D., & Reichardt, C. S. (1991). Random measurement error does not bias the treatment effect estimate in the regression-discontinuity design: I. The case of no interaction. Evaluation Review, 15, 395-419.

Carbonari,J. P., Wirtz, P. W., Muenz, L. R., & Stout, R. L. (1994). Alternative analytical methods for detecting matching effects in treatment outcomes. Journal ofStudies on Alcohol (Suppl. 12), 83-90.

Card, D. (1990). The impact of the Mariel Boatlift on the Miami labor market. Industrial and Labor Relations Review, 43, 245-257.

Carrington, P. J., & Moyer, S. (1994). Gun availability and suicide in Canada: Testing the displacement hypothesis. Studies on Crime and Crime Prevention, 3, 168-178.

Carter, G. M., Winkler, J. D., & Biddle, A. K. (1987). An evaluation of the NIH research career development award. Santa Monica, CA: RAND.

Casella, G., & Schwartz, S.P. (2000). Comment. Journal of the American Statistical Association, 95, 425-428.

Catalano, R., & Serxner, S. (1987). Time series designs of potential interest to epidemiologists. American Journal of Epidemiology, 126, 724-731.

Cecil, J. S., & Boruch, R. F. (1988). Compelled disclosure of research data: An early warning and suggestions for psychologists. Law and Human Behavior, 12, 181-189.

Chaffee, S. H., Roser, C., & Flora, J. (1989). Estimating the magnitude of threats to validity of information campaign effects. In C. G. Salmon (Ed.), Annual review of communication research (Vol. 18). Newbury Park, CA: Sage.

Chalmers, I., Enkin, M., & Keirse, M. J. (Eds.). (1989). Effective care in pregnancy and childbirth. New York: Oxford University Press.

Chalmers, T. C. (1968). Prophylactic treatment of Wilson's disease. New England Journal of Medicine, 278, 910-911.

Chalmers, T. C., Berrier, J., Hewitt, P., Berlin, J., Reitman, D., Nagalingam, R., & Sacks, H. (1988). Meta-analysis of randomized controlled trials as a method of estimating rare complications of non-steroidal anti-inflammatory drug therapy. Alimentary and Pharmacological Therapy, 2-5, 9-26.

Chalmers, T. C., Celano, P., Sacks, H. S., & Smith, H. (1983). Bias in treatment assignment in controlled clinical trials. New England Journal of Medicine, 309, 1358-1361.

Chambless, D. L., & Hollon, S.D. (1998). Defining empirically snppor'ted therapies. Journal of Consulting and Clinical Psychology, 66, 7-18.

Chan, K.-C., & Tumin, J. R. (1997). Evaluating the U.S. nuclear triad. In E. Chelimsky & W R. Shadish (Eds.), Eva/nation for the 21st century: A handbook (pp. 284-298). Thousand Oaks, CA: Sage.

Chaplin, W. F. (1991). The next generation of moderator research in personality psychology. Journal of Personality, 59, 143-178.

Chaplin, W. F. (1997). Personaliry, interactive relations, and applied psychology. In S. R. Briggs, R. Hogan., & W. H. Jones (Eds.), Handbook of personality psychology (pp. 873-890). Orlando, FL: Academic Press.

Chapman, L. J., & Chapman, J, P. (1969). Illusory correlation as an obstacle to the use of valid psychodiagnostic signs. Journal of Abnormal Psychology, 74, 271-280.

Chelimsky, E. (1998). The role of experience in formulating theories of evaluation practice. American Journal of Evalnation, 19, 35-55.

Clien, H., & Rossi, P. H. (1987). The theory-driven approach to validity. Evaluation and Program Planning, 10, 95-103.

Chen, H.-T., & Rossi, P. H. (Eds.). (1992). Using theory to improve program and policy evaluations. New York: Greenwood Press.

Choi, S. C., & Pepple, P. A. (1989). Monitoring clinical trials based on predictive probability of significance. Biometrics, 45, 317-323.

Choi, S.C., Smith, P. J., & Becker, D.P. (1985). Early decision in clinical trials when treatment differences are small: Experience of a controlled trial in head trawna. Controlled Clinical Trials, 6, 280-288.

Ciarlo, J. A., Brown, T. R., Edwards, D. W., Kiresuk, T. J., & Newman, F. L. (1986). Assessing mental health treatment outcome measurement techniques (DHHS Publication No. ADM 86-1301). Washington, DC: U.S. Government Printing Office.

Cicirelli, V. G., and Associates. (1969).The impact of Head Start: An evaluation of the effects of Head Start on children's cognitive and affective development: Vols. 1-2. Athens: Ohio University and Westinghouse Learning Corporation.

Clark, P. I., & Leaverton, P. E. (1994). Scientific and ethical issues in the use of placebo controls in clinical trials. Annual Review of Public Health, 15, 19-38.

Clarridge, B. R., Sheehy, L. L., & Hauser, T. S. (1977). Tracing members of a panel: A 17 year follow-up. InK. F. Schuessler (Ed.), Sociological methodology (pp. 185-203). San Francisco: Jossey-Bass.

Cochran, W. G. (1965). The planning of observational studies in human populations. Journal of the Royal Statistical Society (Series A), 128, 134-155.

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. Biometrics, 24, 295-313.

Cochran, W. G. (1977). Sampling techniques (3rd ed.). New York: Wiley.

Cochran, W. G. (1983). Planning and analysis of observational studies. New York: Wiley.

Cochran, W. G., & Cox, G. M. (1957). Experimental desigus (2nd ed.). New York: Wiley.

Cochran, W. G., & Rubin, D. B. (1973). Controlling bias in observational studies: A, review. Sankhya, 35, 417-446.

Cohen, E., Mowbray, C. T., Bybee, D., Yeich, S., Ribisl, K., & Freddolino, P. P. (1993). Tracking and follow-up methods for research on homelessness. Evaluation Review, 17,331-352.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale NJ: Lawrence Erlbaum Associates.

Cohen, J. (1994). The earth is round (p<.05). American Psychologist, 49, 997-1003.

Cohen, J., & Cohen, P. (1983). Applied multiple regression/correlation analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, P., Cohen, J., Teresi, J., Marchi, M., & Velez, C. N. (1990). Problems in the measurement of latent variables in structural equation causal models. Applied Psychological Measurement, 14, 183-196.

Colditz, G. A., Miller, J. N., & Mosteller, F. (1988). The effect of study design on gain in evaluation of new treatments in medicine and surgery. Drug Information Journal, 22, 343-352.

Collins, H. M. (1981). Stages in the empirical programme of relativism. Social Studies of Science, 11, 3-10.

Collins, J. F., & Elkin, I. (1985). Randomization in the NIMH Treatment of Depression Collaborative Research Program. In R., F. Boruch & W. Wothke (Eds.), Randomization and field experimentation (pp. 27-37). San Francisco: Jossey-Bass.

Comer, J.P. (1988). Educating poor minority children. Scientific American, 259, 42-48.

Connell, D. B., Turner, R. R., & Mason, E. F. (1985). Summary of findings of the school health education evaluation: Health promotion effectiveness, implementation and costs. Journal of School Health, 55, 316-321.

Connell, J. P., Kubisch, A. C., Schorr, L. B., & Weiss, C. H. (Eds.). (1995). New approaches to evaluating community initiatives: Concepts, methods and contexts. Washington, DC: Aspen Institute.

Conner, R. F. (1977). Selecting a control group: An analysis of the randomization process in twelve social reform programs. Evaluation Quarterly, 1, 195-244.

Connor, S. (1989). Postmodernist culture: An introduction to theories ofthe contemporary. Oxford, England: Basil Blackwell.

Connors, A. F., Speroff, T., Dawson, N. V., Thomas, C., Harrell, F. E., Wagner, D., Desbiens, N., Goldman, L., Wu, A. W., Califf, R. M., Fulkerson, W. J ., Vidaillet, H., Broste, S., Bellamy, P., Lynn, J,, & Knaus, W. A. {1996). The effectiveness of right heart catheterization in the initial care of critically ill patients. Journal of the American Medical Asso- ciation, 276, 889-897.

Conrad, K. J. (Ed.). (1994). Critically evaluating the role ofexperiments. San Francisco: Jessey-Bass.

Conrad, K. J., & Conrad, K M. (1994). Reassessing validity threats in experiments: Focus on construct validity. InK. J. Conrad (Ed.), Critically evaluating the role of experiments (pp. 5-25). San Francisco: Jessey-Bass.

Cook, R. D., & Weisberg, S. (1994). An introduction to regression graphics. New York: Wiley.

Cook, T. D. (1984). What have black children gained academically from school integration? Examination of the meta-analytic evidence. InT. D. Cook, D. Armor, R. Crain, N. Miller, W. Stephan, H. Walberg, & P. Wortman (Eds.), School desegregation and black achievement (pp. 6-67). Washington, DC: National Institute of Education. (ERIC Document Reproduction Service No. ED 241 671)

Cook, T. D. (1985). Postpositivist critical multiplism. In L. Shotland & M. M. Mark (Eds.), Social science and social policy (pp. 21-62). Newbury Park, CA: Sage.

Cook, T. D. (1990). The generalization of causal connections: Multiple theories in search of clear practice. In L. Sechrest, E. Perrin, & J. Bunker (Eds.), Research methodology: Strengthening causal interpretations of nonexperimental data (DHHS Publication No. PHS 90-3454, pp. 9-31). Rockville, MD: Department of Health and Hnman Services.

Cook, T. D. (1991). Clarifying the warrant for generalized causal inferences in quasi- experimentation. In M. W. McLaughlin & D. C. Phillips (Eds.), Evaluation and education: At quarter-century (pp. 115-144). Chicago: National Society for the Study of Education.

Cook, T. D. (2000). The false choice between theory-based evaluation and experimentation. In p, J. Rogers, T. A. Hasci, A. Petrosino, & T. A. Huebner (Eds.), Program theory in evaluation: Challenges and opportunities (pp. 27-34). San Francisco, CA: Jossey-Bass.

Cook, T. D., Appleton, H., Conner, R. F., Shaffer, A., Tamkin, G., & Weber, S. J. (1975). «Sesame Street" revisited. New York:' Russell Sage Foundation.

Cook, T. D., Calder, B. J., & Wharton, J.D. (1978). How the introduction oftelevision affected a variety of Social indicators (Vols. 1-4). Arlington, VA: National Science Foundation.

Cook, T.D., & Campbell, D. T. (1979). Quasi-experimentation: Design and analysis issues for field settings. Chicago: Rand·McNally.

Cook, T. D., Cooper, H., Cordray, D. S., Hartmann, H., Hedges, L. V., Light, R. J., Louis, T. A., & Mosteller, F. (Eds.). (1992). Meta-analysis for explanation: A casebook. New York: Russell Sage Foundation.

Cook, T. D., Gruder, C. L., Hennigan, K. M., & Flay, B. R. (1979). History of the sleeper effect: Some logical pitfalls in accepting the null hypothesis. Psychological Bulletin, 86, 662-679.

Cook, T. D., Habib, F. N., Phillips, M., Settersten, R. A., Shagle, S. C., & Degirmencioglu, S. M. (1999). Comer's School Development Program in Prince George's County, Maryland: A theory-based evaluation. American Educational Research Journal, 36, 543-597.

Cook, T. D., Hunt, H. D., & Murphy R. F. (2000). Comer's School Development Program in Chicago: A theory-based evaluation. American Educational Research Journal, 37, 535-597.

Cook, T. D., & Shadish, W. R. (1994). Social experiments: Some developments over the past 15 years. Annual Review of Psychology, 45, 545-580.

Cook, T. D., Shagle, S. C., & Degirmenciogln, S.M. (1997). Capturing social process for testing mediational models of neighborhood effects. Ln J. Brooks-Gunn, G. J. Duncan, & J. L. Aber (Eds.), Neighborhood poverty: Context and consequences for children (Vol. 2). New York: Russell Sage Foundation.

Cooper, H. (1998). Synthesizing research: A guide for literature reviews (3rd ed.). Thousand Oaks, CA: Sage.

Cooper, H., & Hedges, L. V. (Eds.). (1994a). The handbook of research synthesis. New York: Russell Sage Foundation.

Cooper, H., & Hedges, L. V. (1994b). Research synthesis as a scientific enterprise. In H. Cooper & L. V. Hedges (Eds.), The handbook of research synthesis (pp. 3-14). New York: Russell Sage Foundation.

Cooper, W. H., & Richardson, A. J. (1986). Unfair comparisons. Journal of Applied Psychology, 71, 179-184.

Coover, J. E., & Angell, F. (1907). General practice effect of special exercise. American Journal of Psychology, 18,328-340.

Copas, J., & Li, H. (1997). Inference for non-random samples (with discussion). Journal of the Royal Statistical Society (Series B), 59, 55-95.

Cordray, D. S. (1986). Quasi-experimental analysis: A mixture of methods and judgment. In W. M. K. Trochim (Ed.), Advances in quasi-experimental design and analysis (pp. 9-27). San Francisco: Jossey-Bass.

Corrin, W. J., & Cook, T. D. (1998). Design elements of quasi-experimentation. Advances in Educational Productivity, 7, 35-57.

Cosgrove, N., Borhani, N. O., Bailey, G., Borhani, P., Levin, J., Hoffmeier, M., Krieger, S., Lovato, L. C., Petrovitch, H., Vogt, T., Wilson, A. C., Breeson, V., Probstfield, J. L., and the Systolic Hypertension in the Elderly Program (SHEP) Cooperative Research Group. (1999). Mass mailing and staff experience in a total recruitment program for a clinical trial: The SHEP experience. Controlled Clinical.Trials, 19, 133-148.

Costanza, M. C. (1995). Matching. Preventive Medicine, 24, 425-433.

Cowles, M. (1989). Statistics Jn psychology: An historical perspective. Hillsdale, NJ: Erlbaum.

Cox, D. R. (1958). Planning of experiments. New York' Wiley.

Coyle, S. L., Boruch, R. F., & Turner, C. F. (Eds.). (1991). Evaluating AIDS prevention programs (Expanded ed.). Washington, DC: National Academy Press.

Cramer, D. (1990). Self-esteem and close relationships: A statistical refinement. British Journal of Social Psychology, 29, 189-191.

Cramer, J. A., & Spilker, B. (Eds.). (1991). Patient compliance in medical practice and clinical trials. New York: Raven Press.

Critelli, J. W., & Neumann, K. F. (1984). The placebo: Conceptual analysis of a construct in transition. American Psychologist, 39, 32-39.

Crocker, J. (1981). Judgment of covariation by social perceivers. Psychological Bulletin, 90, 272-292,

Cromwell, J. B., Hannan, M. J., Labys, W. C., & Terraza, M. (1994). Multivariate tests for time series models. Thousand Oaks, CA: Sage.

Cromwell, J. B., Labys, W. C., & Terraza, M. (1994). Univariate tests for time series models. Thousand Oaks, CA: Sage.

Cronbach, L. J. (1946). Response sets and test validity. Educational and Psychological Measurement, 6, 475-494.

Cronbach, L. J. (1982). Designing evaluations of educational and social programs. San Francisco: Jossey-Bass.

Cronbach, L. J, (1986). Social inquiry by and for earthlings, In D. W. Fiske & R. A. Shweder (Eds.), Metatheory in social science (pp. 83-107). Chicago: University of Chicago Press. .

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), Test validity (pp. 3-17), Hillsdale, NJ: Erlbaum,

Cronbach, L. J, (1989), Construct validation after thirty years, In R, L. Linn (&L), Intelligence: Measurement, theory and public policy (pp. 147-171). Urbana: University of Illi- nois Press.

Cronbach, L. J., Ambron, S. R., Dornbusch, S. M., Hess, R. D., Hornik, R. C., Phillips, D. C., Walker, D, E, & Weiner S, S. (1980), Toward reform ofprogram evaluation. San Francisco: Jossey-Bass.

Cronbach, L. J., Gieser, G. C., Nanda, H., & Rajaratnam, N, (1972) .The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: Wiley.

Cronbach, L. J., & Meehl, P, E. (1955), Construct validity in psychological tests, Psycho- logiutl Bulletin, 52, 281-302.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C, (1967) .The dependability of behavioral measurements: Multifacet studies of generalizability. Stanford, CA: Stanford University Press.

Cronbach, L J., Rogosa, D, R., Floden, R, K, & Price, G, G, (1977). Analysis of covariance in nonrandomized experiments: Parameters affecting bias (Occasional Paper). Palo Altb, CA: Stanford University, Stanford Evaluation Consortium.

Cronbach, L. J., & Snow, R, E, (1977), Aptitudes and instructional methods: A handbook for research on interactions. New York: lrvingron.

Crosbie, J. (1993). Interrupted time-series analysis with brief single-subject data. Journal of Consulting and Cliniutl Psychology, 61, 966-974.

Cullen, K. W., Koehly, L. M., Anderson, C., Baranowski, T., Prokhorov, A., Basen Engquist, K., Wetter, D., & Hergenroeder, A. (1999). Gender differences in chronic disease risk behaviors through the transition out of high school. American Journal of Preventive Medicine, 17, 1-7.

Cunningham, W. R. (1991). Issues in factorial invariance. In L. M. Collins & J. L. Horn (Eds.), Best methods for the analysis of change: Recent advances, unanswered questions, future directions (pp. 106-113). Washington, DC: American Psychological Association.

Currie, J., & Duncan, T. (1995). Does Head Start make a difference? American Economic Review, 85, 341-364.

Currie, J., & Duncan, T. (1999). Does Head Start help Hispanic children? Journal of Public Economics, 74, 235-262.

D'Agostino, R. B., & Kwan, H. (1995). Measuring effectiveness: What to expect without a randomized control group. Medical Care, 33 (Suppl.), AS95-AS105.

D'Agostino, R. B., & Rubin, D. B. (2000). Estimating and using propensity scores with partially missing data. Journal ofthe American Statistical Association, 95, 749-759.

Dallmayr, F. R., & McCarthy, T. A. (Eds.). (1977). Understanding and social inquiry. Notre Dame, IN: University of Notre Dame Press.

Danziger, K. (1990). Constructing the subject: Historical origins of psychological research. Cambridge, England: Cambridge University Press.

Datta, L.-E. (1997). Multimethod evaluations: Using case studies together with other methods. In E. Chelimsky & W. R. Shadish (Eds.), Evaluation for the 21st century: A handbook (pp. 344-359). Thousand Oaks, CA: Sage.

Davidson, D. (1967). Causal relations. Journal of Philosophy, 64, 691-703.

Davies, P. (1984). Superforce: The search for a grand unified theory of nature. New York: Simon and Schuster.

Davies, P. C. W., & Brown, J. R. (Eds.). (1986). The ghost in the atom? A discussion of the mysteries of quantum physics. Cambridge, England: Cambridge University Press.

Davis, C. E. (1994). Generalizing from clinical trials. Controlled Clinical Trials, 15, 11-14.

Dawid, A. P. (2000). Causal inference without counterfactuals. Journal of the American Statistical Association, 95, 407-448.

Day, N. A., Dunt, D. R., & Day, S. (1995). MaximiZing response to surveys in health program evaluation at minimum cost using multiple methods: Mail, telephone, and visit. Evaluation Review, 19, 436-450.

Dehejia, R. H., & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. Journal of the American Statistical Association, 94, 1053-1062.

Dehue, T. (2000). From deception trials to control reagents: The introduction of the control group about a century ago. American Psychologist, 55,264-268.

DeLeeuw, J., & Kreft, I. (1986). Random coefficient models for multilevel analysis. Journal of Educational Statistics, 11, 57-85.

Della-Piana, G. M. (1981). Film criticism. InN. L. Smith (Ed.), New techniques for evaluation (pp. 274-286). Newbury Park, CA: Sage.

Delucchi, K. L. (1994). Methods for the analysis of binary outcome results in the presence of missing data. Journal of Consulting and Clinical Psychology, 62, 569-575.

Delucchi, K. L., & Bostrom, A. (1999). Small sample lohgitudinal clinical trials with missing data: A comparison of analytic methods. Psychological Methods, 4, 158-172.

Deluse, S. R. (1999). Mandatory divorce education: A program evaluation using a "quasi-random" regression discontinuity design (Doctoral dissertation, Arizona State UniVersity, 1999). Dissertation Abstracts International, 60(03), 1349B.

Dennis, M. L. (1988).Implementing randomized field experiments: An analysis of criminal and civil justice research. Unpublished doctoral dissertation, Northwestern University.

Dennis, M. L., Lennox, R. D., & Foss, M.A. (1997). Practical power analysis for substance abuse health services research. In K. L. Bryant, M. Windell, & S. G. West (Eds.), The science of prevention: Methodological advances from alcohol and substance abuse research (pp. 367-404). Washington, DC: American Psychological Association.

Denton, F. T. (1985). Data mining as an industry. Review of Economics and Statistics, 67, 124-127.

Denton, T. (1994). Kinship, marriage and the family: Eight time series, 35000 B.C. to 2000 A.D. International Journal of Comparative Sociology, 35, 240-251.

Denzin, N. (1989). The research act: A theoretical introduction to sociological methods. Englewood Cliffs, NJ: Prentice-Hall.

Denzin, N. K., & Lincoln, Y. S. (2000). Handbook of qualitative research (2nd ed.). Newbury Park, CA: Sage.

Devine, E. C. (1992). Effects of psychoeducational care with adult surgical patients: A theory-probing meta-analysis of intervention studies. In T. Cook, H. Cooper, D. Cordray, H. Hartmann, L. Hedges, R. Light, T. Louis, & F. Mosteller (Eds.), Meta-analysis for explanation: A casebook (pp. 35-82). New York: Russell Sage Foundation.

Devine, E. C., & Cook, T. D. (1983). A meta-analytic analysis of effects of psychoeducational interventions on length of post-surgical hospital stay. Nursing Research, 32, 267-274.

Devine, E. C., & Cook, T. D. (1986). Clinical and cost-saving effects of psychoeducational interventions with surgical patients: A meta-analysis. Research in Nursing and Health, 9, 89-105.

Devine, E. C., O'Connor, F. W., Cook, T. D., Wenk, V. A., & Curtin, T. R. (1988). Clinical and financial effects of psychoeducational care provided by staff nurses to adult surgical patients in the post-DRG environment. American Journal of Public Health, 78, 1293-1297.

Devlin, B. (Ed.). (1997). Intelligence and success. Is it all in the genes? Scientists respond to The Bell Curve. New York: Springer-Verlag.

Diament, C., & Colletti, G. (1978). Evaluation of behavioral group counseling for parents of learning-disabled children. Journal of Abnormal Child Psychology, 6, 385-400.

Diaz-Guerrero, R., & Holtzman, W. H. (1974). Learning by televised "Plaza Sesamo" in Mexico. Journal of Educational Psychology, 66, 632-643.

Dickerson, K. (1994). Research registers. In H. Cooper & L. V. Hedges (Eds.), The handbook of research synthesis (pp. 71-83). New York: Russell Sage Foundation.

Dickerson, K., Higgins, K., & Meinert, C. L. (1990). Identification of meta-analyses: The need for standard terminology. Controlled Clinical Trials, 11, 52-66.

Diehr, P., Martin, D. C., Koepsell, T., & Cheadle, A. (1995). Breaking the matches in a paired t-test for community interventions when the number of pairs is small. Statistics in Medicine, 14, 1491-1504.

DiRaddo, J.D. (1996). The investigation and amelioration of a staff turnover problem. Dissertation Abstracts International, 59 (04), 1133A (University Microfilms No. 316379).

Director, S. M. (1979). Underadjustment bias in the evaluation of manpower training. Evaluation Review, 3, 190-218.

Dixon, D. O., & Lagakos, S. W. (2000). Should data and safety monitoring boards share confidential interim data? Controlled Clinical Trials, 21, 1-6.

Dohrenwend, B. P., Shrout, P. E., Egri, G., & Mendelsohn, F. S. (1980). Nonspecific psychological distress and other dimensions of psychopathology. Archives of General Psychiatry, 37, 1229-1236.

Donner, A. (1992). Sample size requirements for stratified cluster randomization designs. Statistics in Medicine, 11, 743-750.

Donner, A., & Klar, N. (1994). Cluster randomization trials in epidemiology: Theory and application. Journal of Statistical Planning and Inference, 42, 37-56.

Donner, A., & Klar, N. (2000). Design and analysis of cluster randomization trials in health research. London: Arnold.

Drake, C. (1993). Effects of misspecification of the propensity score on estimators of treatment effects. Biometrics, 49, 1231-1236.

Drake, C., & Fisher, L. (1995). Prognostic models and the propensity score. International Journal of Epidemiology, 24, 183-187.

Drake, S. (1981). Cause, experiment, and science. Chicago: University of Chicago Press.

Draper, D. (1995). Inference and hierarchical modeling in social sciences. Journal of Educational and Behavioral Statistics, 20, 115-147.

Droitcour, J. A. (1997). Cross-design synthesis: Concepts and applications. In E. Chelimsky & W. R. Shadish (Eds.), Evaluation for the 21st century: A handbook (pp. 360-372).Thousand Oaks, CA: Sage.

Ducasse, C. J. (1951). Nature, mind and death. La Salle, IL: Open Court.

Duckart, J.P. (1998). An evaluation of the Baltimore Community Lead Education and Reduction Corps (CLEARCorps) Program. Evaluation Review, 22, 373-402.

Dukes, R. L., Ullman, J. B., & Stein, J. A. (1995). An evaluation of D.A.R.E. (Drug Abuse Resistance Education), using a Solomon Four-Group design with latent variables. Evaluation Review, 19, 409-435.

Duncan, G. J., Yeung, W. J., Brooks-Gunn, J., & Smith, J. R. (1998). How much does childhood poverty affect the life chances of children? American Sociological Review, 63, 406-423.

Dunford, F. W. (1990). Random assignment: Practical considerations from field experiments. Evaluation and Program Planning, 13, 125-132.

Durlak, J. A., & Lipsey, M. W. (1991). A practitioner's guide to meta-analysis. American Journal of Community Psychology, 19, 291-332.

Duval, S., & Tweedie, R. (2000). A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. Journal of the American Statistical Association, 95, 89-98.

Dwyer, J. H., & Flesch-Janys, D. (1995). Agent Orange in Vietnam. American Journal of Public Health, 85, 476-478.

Edgington, E. S. (1987). Randomized single-subject experimentS and statistical tests. Journal of Counseling Psychology, 34, 437-442.

Edgington, E. S. (1992). Nonparametric tests for single-case experiments. In T. R. Kratochwill & J. R. Levin (Eds.), Single-case research design and analysis (pp. 133-157).Hillsdale, NJ: Erlbaum.

Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. Psychological Methods, 5, 155-174.

Eells, E. (1991). Probabilistic causality. New York: Cambridge University Press.

Efron, B. (1971). Forcing a sequential experiment to be balanced. Biometrika, 58, 403-417.

Efron, B., & Feldman, D. (1991). Compliance as an explanatory variable in clinical trials. Journal of the American Statistical Association, 86, 9-26.

Efron, B., & Tibshiran, R. J. (1993). An introduction to the bootstrap. New York: Chapman & Hall.

Eggan, F. (1954). Social anthropology and the method of controlled comparison. American Anthropologist, 56, 743-763.

Einstein, A. (1949). Reply to criticisms. In P. A. Schilpp (Ed.), Albert Einstein: Philosopherscientist (pp. 665-688). Evanston, IL: Library of Living Philosophers.

Eisenhart, M., & Howe, K. (1992). Validity in educational research. In M.D. LeCompte, W. L. Millroy, & J. Preissle (Eds.), The handbook of qualitative research in education (pp. 643-680). San Diego: Academic Press.

Eisner, E. (1979). The educational imagination. New York: Macmillan.

Eisner, E. (1983). Anastasia might still be alive, but the monarchy is dead. Educational Researcher, 12, 5.

Elboume, D., Garcia, J., & Snowdon, C. (1999). Reply. Controlled Clinical Trials, 20,571-572.

Elkin, l., Parloff, M. B., Hadley, S. W., & Autry, J. H. (1985). NIMH Treatment of Depression Collaborative Research Program: Background and research plan. Archives of General Psychiatry, 42, 305-316.

Elkin, I., Shea, T., Watkins, J. T., Imber, S. D., Sotsky, S. M., Collins, J. F., Glass, D. R., Pilkonis, P. A., Leber, W. R., Docherty, J. P., Fiester, S. J., & Parloff, M. B. (1989). National Institute of Mental Health Treatment of Depression Collaborative Research Program: General effectiveness of treatments. Archives of General Psychiatry, 46, 971-982.

Ellenberg, J. H. (1994). Cohort studies: Selection bias in observational and experimental studies. Statistics in Medicine, 13, 557-567.

Ellenberg, S. S., Finkelstein, D. M., & Schoenfeld, D. A. (1992). Statistical issues arising in AIDS clinical trials. Journal of the American Statistical Association, 87, 562-569.

Elmore, R. F. (1996). Getting to scale with good educational practice. Harvard Educational Review, 66, 1-26.

Emanuel, E. J., Wendler, D., & Grady, C. (2000). What makes clinical research ethical? Journal of the American Medical Association, 283, 2701-2711.

Embretson, S. (1983). Construct validity: Construct representation versus nomothetic span. American Psychologist, 93, 179-197.

Emerson, R. M. (1981). Observational field work. Annual Review of Sociology, 7, 351-378.

Emerson, S. S. (1996). Statistical packages for group sequential methods. American Statistician, 50, 183-192.

Epperson, D. L., Bushway, D. J., & Warman, R. E. (1983). Client self-termination after one counseling session: Effects of problem recognition, counselor gender, and counselor experience. Journal of Counseling Psychology, 30, 307-315.

Equal Employment Opportunity Commission, Departmeut of Labor, Department of Justice, and the Civil Service Commission. (1978, August). Adoption by four agencies of uniform guidelines on employee selection procedures. 34 Fed. Reg. 38290-38315.

Erbland, M. L., Deupree, R. H., & Niewoehner, D. E. (1999). Systemic corticosteroids in chronic obstructive pulmonary disease exacerbations (SCCOPE): Rationale and design of an equivalence trial. Controlled Clinical Trials, 19,404-417.

Erez, E. (1986). Randomized experiments in correctional context: Legal, ethical, and practical concerns. Journal of Criminal Justice, 14, 389,400.

Esb nsen, F.-A., Deschenes, E. P., Vogel, R. E., West, J., Arboit, K., & Harris, L. (1996). Active parental consent in school-based research: An examination of ethical and methodological issues. Evaluation Review, 20, 737-753.

Estes, W. K. (1997). Significance testing in psychological research: Some persisting issues. Psychological Science, 8, 18-20.

Estroff, S. E. (1981). Making it crazy: An ethnography of psychiatric clients in an American community. Berkeley: University of California Press.

Etzioni, R. D., & Kadane, J. B. (1995). Bayesian statistical methods in public health and medicine. Annual Review of Public Health, 16, 23-41.

Everitt, D. E., Soumerai, S. B., Avorn, J., Klapholz, H., & Wessels, M. (1990). Changing surgical antimicrobial prophylaxis practices through education targeted at senior department leaders. Infectious Control and Hospital Epidemiology, 11, 578-583.

Eyberg, S. M., & Johnson, S.M. (1974). Multiple assessment of behavior modification with families: Effects of contingency contracting and order of treated problems. Journal of Consulting and Clinical Psychology, 42, 594-606.

Eysenck, H. J. (1978). An exercise in mega-silliness. American Psychologist, 33, 517.

Eysenck, H. J., & Eysenck, M. (1983). Mindwatching: Why people behave the way they do. Garden City, NY: Anchor Press.

Fagan, J. A. (1990). Natural experiments in criminal justice. In K. L. Kempf (Ed.), Measurement issues in criminology {pp. 108-137). New York: Springer-Verlag.

Fagerstrom, D. 0. (1978). Measuring degree of physical dependence to tobacco smoking with reference to individualization of treatment. Addictive Behaviors, 3, 235-241.

Fairweather, G. W., & Tornatsky, L. G. (1977). Experimental methods for social policy research. New York: Pergamon Press.

Faith, M.S., Allison, D. B., & Gorman, B.S. (1997). Meta-analysis of single-case research. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), Design and analysis of single-case research (pp. 245-277). Hillsdale, NJ: Erlbaum.

Family Support Act, Pub. L. N. 100-485, Section 203, 102 Stat. 2380 (1988).

Farquhar, J. W., Fortmann, S. P., Flora, J. A., Taylor, C. B., Haskell, W. L., Williams, P. T., MacCoby, N., & Wood, P. D. (1990). The Stanford five-city project: Effects of community-wide education on cardiovascular disease risk factors. Journal of the American Medical Association, 26, 359-365.

Faust, D. (1984). The limits of scientific reasoning. Minneapolis: University of Minnesota Press.

Federal Judicial Center. (1981). Experimentation in the law: Report of the Federal Judicial Center Advisory Committee on Experimentation in the Law. Washington, DC: U.S. Government Printing Office.

Feinauer, D. M., & Havlovic, S. J. (1993). Drug testing as a strategy to reduce occupational accidents: A longitudinal analysis.Journal of Safety Research, 24, 1-7.

Feinberg, S. E. (1971). Randomization and social affairs: The 1970 drah lottery. Science, 171, 255-261.

Feinberg, S. E., Singer, B., & Tanur, J. M. (1985). Large-scale social experitnentation in the United States. In A, C. Atkinson & S. E. Feinberg (Eds.), A celebration ofstatistics: The ISI centenary volume (pp. 287-326). New York: Springer-Verlag.

Feldman, H. A., & McKinlay, S.M. (1994). Cohort versus cross-sectional design in large field trials: Precision, sample size, and a unifying model. Statistics in Medicine, 13, 61-78.

Feldman, H. A., McKinlay, S. M., & Niknian, M. (1996). Batch sampling to improve power in a community trial: Experience from the Pawtucket Heart Health Program. Evaluation Review, 20, 244-274.

Feldman, R. (1968). Response to compatriot and foreigner who seek assistance. Journal of Personality and Social Psychology, 10, 202-214.

Festinger, L. {1953). Laboratory experiments. In L. Festinger & D. Katz (Eds.), Research methods in the behavioral sciences (pp. 136-172). New York: Holt, Rinehart & Winston.

Fetterman, b. M. (1982). Ibsen's baths: Reactivity and insensitivity. Educational Evaluation and Policy Analysis, 4, 261-279.

Fetterrilah, D. M. (Ed.). (1984). Ethnography in educational evaluation. Beverly Hills, CA:

Sage.

Feyeiabend, P. (1975). Against method: Outline of an anarchisitic theory of knowledge. Atlantic Highlands, NJ: Humanities Press.

Feyefabend, P. (1978). Science in a free society. London: New Left Books.

Filstead, W. (1979). Qualitative methods: A needed perspective in evaluation research. In T. Cook & C. Reichardt (Eds.), Qualitative and quantitative methods in evaluation research (pp. 33-48). Newbiiry Park, CA: Sage.

Fink, A. (1998). Conducting research literature reviews. Thousand Oaks, CA: Sage.

Firlkelstein, M. O., Levin, B., & Robbins, H. (1996a). Clinical and prophylactic trials with assured new treatment for those at greater risk: I. A design proposal. American Journal of Public Health, 86, 691-695.

Finkelstein, M. O., Levin, B., & Robbins, H. (1996b). Clinical and prophylactic trials with assured new treatment for those at greater risk: Il Examples. American Journal of Publit Health, 86, 696-705.

Finn, J.D., & Achilles, C.M. (1990). Answers and questions about class size: A statewide experiment. American Educational Research Journal, 27, 557-577.

Fischer, R. (1994). Control construct design in evaluating campaigns. Public Relations Review, 21, 45-58.

Fischer-Lapp, K., & Goetghebeur, E. (1999). Practical properties of some structural mean analyses of the effect of compliance in randomized trials. Controlled Clinical Trials, 20, 531-546.

Fischhoff, B. (1975). Hindsight/foresight: The effect of outcome knowledge on judgment under uncertainty. Journal of Experimental Psychology: Human Perception and Performance, 1, 288-299.

Fisher, L. D. (1999). Advances in clinical trials in the twentieth century. Annual Review of Public Health, 20, 109-124.

Fisher, R. A. (1925). Statistical methods for research workers. Edinburgh: Oliver & Boyd.

Fisher, R. A. (1926). The arrangement of field experiments.Journal of the Ministry of Agriculture of Great Britain, 33,505-513.

Fisher, R. A. (1932). Statistical methods for research workers (4th ed.). London: Oliver & Boyd.

Fisher, R. A. (1935). The design of experiments. Edinburgh: Oliver & Boyd.

Fisher, R. A., & Yates, F. (1953). Statistical tables for biological, agricultural, and medical research (4th ed.). Edinburgh: Oliver & Boyd.

Fleiss, J. L. (1981). Statistical methods for rates and proportions (2nd ed.). New York: Wiley.

Fleiss, J. L. (1986). The design and analysis of clinical experiments. New York: Wiley.

Fleiss, J. L. (1994). Measures of effect size for categorical data. In H. Cooper &: L. V. Hedges (Eds.), The handbook of research synthesis (pp. 245-260). New York: Russell Sage Foundation.

Flournoy, N., & Rosenbergei; W. F. (Eds.). (1995). Adaptive designs. Hayward, CA: IMS.

Folkman, J. (1996). Fighting cancer by attacking its blood supply. Scientific American, 275,150-154.

Fortin, F., & Kirouac, S. (1976). A randomized controlled trial of preoperative patient education. International Journal of Nursing Studies, 13, 11-24.

Foster, E. M., & Bickman, L. (1996). An evaluator's guide to detecting attrition problems. Evaluation Review, 20, 695-723.

Foster, E. M., & McLanahan, S. (1996). An illustration of the use of instrumental variables: Do neighborhood conditions affect a young person's chance of finishing high school? Psychological Methods, 1, 249-261.

Fowler, R. L. (1985). Testing for substantive significance in applied research by specifying nonzero effect null hypotheses. Journal of Applied Psychology, 70, 215-218.

Fraker, T., & Maynard, R. {1986, October). The adequacy of comparison group designs for evaluations of employment-related programs. (Available from Mathematica Policy Research, P.O. Box 2393, Princeton, NJ 08543-2393)

Fraker, T., & Maynard, R. (1987). Evaluating comparison group designs with employmentrelated programs. Journal of Human Resources, 22, 194-227.

Frangakis, C. E., & Rubin, D. B. (1999) Addressing complications of intention-to-treat analysis in the combined presence of ali-or-none treatment-noncompliance and subsequent missing outcomes. Biometrika, 86, 366-379.

Frankel, M. (1983). Sampling theory. In P. H. Rossi, J. D. Wright, & A. B. Anderson (Eds.), Handbook ofsurvey research (pp. 21-67). San Diego: Academic Press.

Franklin, C., Grant, D., Corcoran, J., Miller, P. O., & Bultman, L. (1997). Effectiveness of prevention programs for adolescent pregnancy: A meta-analysis. Journal of Mariage and the Family, 59,551-567.

Franklin, R. D., Allison, D. B., & Gorman, B. S. (Eds.). (1997). Design and analysis of single-case research. Mahwah, NJ: Erlbaum.

Freedman, D. A. (1987). As others see us: A case study in path analysis. Journal of Educational Statistics, 12, 101-128.

Franklin, R. D., Allison, D. B., & Gorman, B. S. (Eds.). (1997). Design and analysis of single-case research. Mahwah, NJ: Erlbawn.

Freedman, D. A (1987). As others see us: A case study in path analysis. Journal of Educational Statistics, 12, 101-128.

Freedman, L. S., & White, S. J. (1976). On the use of Pocock and Simon's method for balancing treatment numbers over prognostic variables in the controlled clinical trial. Biometrics, 32, 691-694.

Freiman, J. A., Chalmers, T. C., Smith, H., & Kuebler, R. R. (1978). The importance of beta, the Type II error, and sample size in the design and interpretation of the randomized control trial. New England Journal of Medicine, 299, 690-694.

Frick, R. W. (1995). Accepting the null hypothesis. Memory & Cognition, 23, 132-138.

Frick, R. W. (1996). The appropriate use of null hypothesis testing. Psychological Methods, 1, 379-390.

Friedlander, D., & Robins, P. K. (1995). Evaluating program evaluations: New evidence on commonly used nonexperimental methods. American Economic Review, 85, 923-937.

Friedman, J., & Weinberg, D. H. (Eds.). (1983). Urban affairs annual review: Volume 24, The great housing experiihent. Thousand Oaks, CA: Sage.

Fuller, H. (2000). Evidence supports the expansion of the Milwaukee parental choice program. Phi Delta Kappan, 81, 390-391.

Fuller, W. A. (1995). Introduction to statistical time series (2nd ed.). New York: Wiley.

Furby, L. {1973). Interpreting regression toward the mean in development research. Developmental Psychology, 8, 172-179.

Furlong, M. J., Casas, J. M., Corrall, C., & Gordon, M. (1997). Changes in substance use patterns associated tvith the development of a community partnership project. Evaluation and Program Planning, 20, 299-305.

Furlong, M. J., & Wampold, B. E. (1981). Visual analysis of single-subject studies by school psychologists. Psychology in the Schools, 18, 80-86.

Gadenne, V. (1976). Die Gultigkeit psychologischer Unterscuchungen. Stuttgart, Germany: Kohlhammer.

Gail, M. H., Byar, D. P., Pechacek, T. F., & Corle, D. K. {1992). Aspects of statistical design for the Community Intervention Trial for Smoking Cessation (COMMIT). Controlled Clinical Trials, 13, 6-21.

Gail, M. H., Mark, S.D., Carroll, R. J., Green, S. B., & Pee, D. (1996). On desigrt considerations and randomization-based inference for community intervention trials. Statistics in Medicine, 15, 1069-1092.

Gallini, J. K., & Bell, M. E. (1983). Formulation of a structural equation model for the evaluation of curriculum. Educational Evaluation and Policy Analysis, 5, 319-326.

Galton, F. (1872). Statistical inquiries into the efficacy of prayer. Fortnightly Review, 12, 124-135.

Galton, F. (1886). Regression towards mediocrity in hereditary stature. Journal of the Anthropological Institute, 15, 246-263.

Garber, J., & Hollon, S. D. (1991).What can specificity designs say about causality in psychopathology research? Psychological Bulletin, 110, 129-136.

Gastwirth, J. (1992). Method for assessing the sensitivity of statistical comparisons used in Title Vll cases to omitted variables. Jurimetrics, 33, 19-34.

Gastwirth, J., Krieger, A., & Rosenbaum, P. (1994). How a court accepted an impossible explanation. American Statistician, 48, 313-315.

Geertz, C. (1973). Thick description: Toward an interpretative theory of culture. In C. Geertz (Ed.), The interpretation of culture (pp. 3-30). New York: Basic Books.

Gephart, W. J. (1981). Watercolor painting. In N. L. Smith (Ed.), New techniques for evaluation (pp. 286-298). Newbury Park, CA: Sage.

Gergen, K. J. (1973). Social psychology as history. Journal of Personality and Social Psychology, 26, 309-320.

Gerdnimus, A. T., & Korenman, S. (1992). The socioeconomic consequences of teen childbearing reconsidered. Quarterly Journal of Economics, 107, 1187-1214.

Gholson, B., & Houts, A. C. (1989). Toward a cognitive psychology of science. Social Epistemology, 3, 107-127.

Gholson, B. G., Shadish, W. R., Neimeyer, R. A., & Houts, A. C. (Eds.). (1989). Psychology of science: Contributions to metascience. Cambridge, Eilgland: Cambridge University Ptess.

Gibbons, R. D., Redeker, D. R., & Davis, J. M. (1993). Estimation of effect size from a series of experiments involving paired comparisons. Journal of Educational Statistics, 18, 271-279.

Gigerenzer, G. (1996). On narrow norms and vague heuristics: A reply to Kahneman and Tversky. Psychological Review, 103, 592-596.

Gilbert, J.P., McPeek, B., & Mosteller, F. (1977a). Progress in surgery and anesthesia: Benefits and risks of innovative therapy. In J.P. Bunker, B. A Barnes, & F. Mosteller (Eds.), Costs, risks, and benefits of surgery (pp. 124-169). New York: Oxford University Press.

Gilbert, J.P., McPeek, B., & Mosteller, F. (1977b). Statistics and ethics in surgery and anesthesia. Science, 198, 684-689.

Gillespie, R. (1988). The Hawthorne experiments and the politics of experimentation. In J. Morawski (Ed.), The rise of experimentation in American psychology (pp. 114-137). New Haven, CT: Yale University Press.

Gilovich, T. (1991). How we know what isn't so: The fallibility of human reasoning in everyday life. New York: Free Press.

Glaser, B. G., & Strauss, A. L. (1967). The discovery of grounded theory: Strategies for qualitative research. New York: Aldine.

Glasgow, R. E., Vogt, T. M., & Boles, S.M. (1999). Evaluating the public health impact of health promotion interventions: The RE-AIM framework. American Journal of Public Health, 89, 1322-1327.

Glass, G. V. (1976). Primary, secondary, and meta-analysis. Educational Researcher, 5, 3-8.

Glass, G. V., & Smith, M. L. (1979). Meta-analysis of research on the relationship of class-size and achievement. Educational Evaluation and Policy Analysis, 1, 2-16.

Gleser, L. J., & Olkin, I. (1994). Stochastically dependent effect sizes. In H. Cooper & L. V. Hedges (Eds.), The handbook of research synthesis (pp. 339-355). New York: Russell Sage Foundation.

Glymour, C., Scheines, R., Spirtes, P., & Kelly, K. (1987). Discovering causal structure: Artificial intelligence, philosophy of science, and statistical modeling. San Diego, CA: Aca- demic Press.

Goetghebeur, E., & Molenberghs, G. (1996). Causal inference in a placebo-controlled clinical trial with binary outcome and ordered compliance. Journal of the American Statistical Association, 91, 928-934.

Goetghebeur, E., & Shapiro, S. H. (1996). Analyzing non-compliance in clinical trials: Ethical imperative or mission impossible? Statistics in Medicine, 15, 2813-2826.

Goetz, J.P., & LeCompte, M.D. (1984). Ethnography and qualitative design in educational research. San Diego, CA: Academic Press.

Goldberg, H. B. (1997, February). Prospective payment in action: The National Home Health Agency demonstration. CARING, 17(2), 14-27.

Goldberger, A. S. (1972a). Selection bias in evaluating treatment effects: Some formal illustrations (Discussion Paper No. 123). Madison: University of Wisconsin, Institute for Research on Poverty.

Goldberge, A. S. (1972b). Selection bias in evaluating treatment effects: The case of interaction (Discussion paper). Madison: University of Wisconsin, Institute for Research on Poverty.

Goldman, J. (1977). A randomization procedure for "trickle-process" evaluations. Evaluation Quarterly, 1, 493-498.

Goldschmidt, W. (1982). [Letter to the editor]. American Anthropologist, 84, 641-643.

Goldstein, H. (1987). Multilevel models in educational and social research. London: Oxford University Press.

Goldstein, H., Yang, M., Omar, R., Turner, R., & Thompson, S. (2000). Meta-analysis using multilevel models with an application to the study of class size effects. Applied Statistics, 49, 399-412.

Goldstein, J. P. (1986). The effect of motorcycle helmet use on the probability of fatality and the severity of head and neck injuries. Evaluation Review, 10, 355-375.

Gooding, D., Pinch, T., & Schaffer, S. (1989b). Preface. In D. Gooding, T. Pinch, & S. Schaffer (Eds.), The uses of experiment: Studies in the natural sciences (pp. xiii-xvii). Cambridge, England: Cambridge University Press.

Goodman, J. S., & Blum, T. C. (1996). Assessing the non-random sampling effects of subject attrition in longitudinal research. Journal of Management, 22, 627-652.

Goodson, B. D., Layzer, J. I., St. Pierre, R. G., Bernstein, L. S. & Lopez, M. (2000). Effectiveness of a comprehensive five-year family support program on low-income children and their families: Findings from the Comprehensive Child Development Program. Early Childhood Research Quarterly, 15, 5-39.

Gorman, B. S., & Allison, D. B. {1997). Statistical alternatives for single-case designs. In R. D. Franklin, D. B. Allison, & B.S. Gorman (Eds.), Design and analysis of single-case research (pp. 159-214). Hillsdale, NJ: Erlbaum.

Gorman, M. E. (1994). Toward an experimental social psychology of science: Preliminary results and reflexive observations. In W. R. Shadish & S. Fuller (Eds.), The social psychology of science (pp. 181-196). New York: Guilford Press.

Gosnell, H. F. (1927). Getting out the vote. Chicago: University of Chicago Press.

Graham, J. W., & Donaldson, S.l. (1993). Evaluating interventions with differential attrition: The importance of nonresponse mechanisms and use of follow-up data. Journal of Applied Psychology, 78, 119-128.

Grandy, J. (1987). Characteristics of examinees who leave questions unanswered on the GRE general test under rights-only scoring (GRE Board Professional Rep. No. 83-16P; ETS Research Rep. No. 87-38). Princeton, NJ: Educational Testing Service.

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. Econometrica, 37, 424-438.

Green, S. B., Corle, D. K., Gail, M. H., Mark, S.D., Pee, D., Freedman, L. S., Graubard, B. I., & Lynn, W. R. (1995). Interplay between design and analysis for behavioral intervention trials with community as the unit of randomization. American Journal of Epidemiology, 142, 587-593.

Greenberg, D., & Shrader, M. (1997). The digest of social experiments (2nd ed.). Washington, DC: Urban Institute Press.

Greenberg, J., & Folger, R. (1988). Controversial issues in social research methods. New York: Springer·Verlag.

Greenberg, R. P., Bornsteffi, R. F., Greenberg, M.D., & Fisher, S. (1992). A meta-analysis of antidepressant outcome under "blinder" conditions. Journal of Consulting and Clinical Psychology, 60, 664-669.

Greene, C. N., & Podsakoff, P.M. (1978). Effects of removal of a pay incentive: A field experiment. Proceedings of the Academy of Management, 38,206-210.

Greene, J.P., Peterson, P. E., & Du, J. (1999). Effectiveness of school choice: The Milwaukee experiment. Education and Urban Society, 31, 190-213.

Greene, W. H. (1985). LIMDEP: An econometric modeling program for the IBM PC.American Statistician, 39, 210.

Greene, W. H. (1999). Econometric analysis. Upper Saddle River, NJ: Prentice·Hall.

Greenhouse, J. B., & Iyengar, S. (1994). Sensitivity analysis and diagnostics. In H. Cooper & L. V. Hedges (Eds.), The handbook of research synthesis (pp. 383-398). New York: Russell Sage Foundation.

Greenhouse, S. W (1982). Jerome Cornfield's contributions to epidemiology. Biometrics, 28 (Suppl.), 33-45.

Greenland, S., & Robins, J. M. (1986). Identifiability, exchangeability, and epidemiological confounding. International Journal of Epidemiology, 15, 413-419.

Greenwald, A. G. {1975). Consequences of prejudice against the null hypothesis. Psychological Bulletin, 82, 1-20.

Greenwald, A. G., Gonzalez, R., Harris, R. J., & Guthrie, D. (1996). Effect sizes and p values: What should be reported and what should be replicated? Psychophysiology, 33, 175-183.

Greenwald, P., & Cullen, J. W. (1984). The scientific approach to cancer control. CA·A Cancer journal for Clinicians, 34, 328-332.

Greenwood, J. D. (1989). Explanation and experiment in social psychological science: Realism and the social constitution of action. New York: Springer-Verlag.

Griffin, L., & Ragin, C. C. (Eds.). (1994). Formal methods of qualitative analysis [Special issue]. Sociological Methods and Research, 23(1).

Grilli, R., Freemantle, N., Minozzi, S., Domenighetti, G., & Finer, D. (2000). Mass media interventions: Effects on health services utilization (Cochrane Review). The Cochrane Library, Issue 3. Oxford, England: Update Software.

Gross, A. J. (1993). Does exposure to second-hand smoke increase lung cancer risk? Chance: New Directions for Statistics and Computing, 6, 11-14.

Grossarth-Maticek, R., & Eysenck, H. J. (1989). Is media information that smoking causes illness a self-fulfilling prophecy? Psychological Reports, 65, 177-178.

Grossman, J., & Tierney, J. P. (1993). The fallibility of comparison groups. Evaluation Review, 17, 556-571.

Groves, R. M. (1989). Survey errors and survey costs. New York: Wiley.

Gu, X. S., & Rosenbaum, P.R. (1993). Comparison of multivariate matching methods: Structures, distances, and algorithms. Journal of Computational and Graphical Statistics, 2, 405-420.

Guba, E., & Lincoln, Y. (1982). Effective evaluation. San Francisco: Jossey-Bass.

Guba, E., & Lincoln, Y. (1989). Fourth generation evaluation. Newbury Park, CA: Sage.

Guba, E. G. (1981). Investigative journalism. In N. L. Smith (Ed.), New techniques for evaluation (pp. 167-262). Newbury Park, CA: Sage.

Guba, E. G. (Ed.). (1990). The paradigm dialog. Newbury Park, CA: Sage.

Gueron, J. M. (1985). The demonstration of state work/welfare initiatives. In R. F. Boruch & W. Wothke (Eds.), Randomization and field experimentation (pp. 5-13). San Francisco: Jossey-Bass.

Gueron, J. M. (1999, May). The politics of random assignment: Implementing studies and impacting policy. Paper presented at the conference on Evaluation and Social Policy in Education of the American Academy of Arts and Sciences, Cambridge, MA. Guilford, J.P. (1946). New standards for test evaluation. Educational and Psychological Measurement, 6, 427-439.

Gunn, W.J., Iverson, D. C., & Katz, M. (1985). Design of school health education evaluation. Journal of School Health, 55, 301-304.

Gurman, A. S., & Kniskern, D. P. (1978). Research on marital and family therapy: Progress, perspective, and prospect. In S. L. Garfield & A. E. Bergin (Eds.), Handbook of psychotherapy and behavior change: An empirical analysis (2nd ed., pp. 817-901). New York: Wiley.

Gwadz, M., & Rotheram-Borus, M. J. (1992, Fall). Tracking high-risk adolescents longitudinally. AIDS Education and Prevention (Suppl.), 69-82.

Haavelmo, T. (1944, July). The probability approach in econometrics. Econometrica, 12 (Suppl.).

Hacking, I. (1983). Representing and interoening: Introductory topics in the philosophy of natural science. Cambridge, England: Cambridge University Press.

Hacking, I. (1988). Telepathy: Origins of randomization in experimental design. Isis, 79, 427-451.

Hackman, J. R., Pearce, J. L., & Wolfe, J. C. (1978). Effects of changes in job characteristics on work attitudes and behaviors: A naturally occurring quasi-experiment. Organizational Behavior and Human Performance, 21, 289-304.

Haddock, C. K., Rindskopf, D., & Shadish, W. R. (1998). Using odds ratios as effect sizes for meta-analysis of dichotomous data: A primer on methods and issues. Psychological Methods, 3, 339-353.

Haddock, C. K., Shadish, W. R., Klesges, R. C., & Stein, R. J. (1994). Treatments for childhood and adolescent obesity: A meta-analysis. Annals of Behavioral Medicine, 16, 235-244.

Hahn, G. J. (1984). Experimental design in the complex world. Technometrics, 26, 19-31.

Halvorsen, K. T. (1994). The reporting format. In H. Cooper & L. V. Hedges (Eds.), The handbook of research synthesis (pp. 425-437). New York: Russell Sage Foundation.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Thousand Oaks, CA:: Sage.

Hamilton, J.D. (1994). Time series analysis. Princeton, NJ: Princeton University Press.

Hand, H. H., & Slocum, J. W., Jr. (1972). A longitudinal study of the effects of a human relations training program on managerial effectiveness. Journal of Applied PsycholOgy, 56,412-417.

Hankin, J. R., Sloan, J. J., Firestone, I. J., Ager, J. W., Sokol, R. J., & Martier, S. S. (1993). A time series analysis of the impact of the alcohol warning label on antenatal drinking. Alcoholism: Clinical and Experimental Research, 17, 284-289.

Hannan, E. G., & Deistler, M. (1988). The statistical theory of linear systems. New York: Wiley.

Hannan, P. J., & Murray, D. M. (1996). Gauss or Bernoulli? A Monte Carlo comparison of the performance of the linear mixed-model and the logistic mixed-model analyses in simulated community trials with a dichotomous outcome variable at the individual level. Evaluation Review, 20, 338-352.

Hansen, M. H., & Hurwitz, W. N. (1996, March). The problem of non-response in sample surveys. Amstat News, 25-26.

Hansen, M. H., Hurwitz, W. N., & Madow, W. G. (1993). Sample survey methods and theory (Vols. 1-2). Somerset, NJ: Wiley.

Hansen, W. B., Tobler, N. S., & Graham, J. W. (1990). Attrition in substance abuse prevention research: A meta-analysis of 85 longitudinally followed cohorts. Evaluation Review, 14, 677-685.

Hanson, N. R. (1958). Patterns of discovery: An inquiry into the conceptual foundations of science. Cambridge, England: Cambridge University Press.

Hanushek, E. A. (1999). The evidence on class size. InS. E. Mayer & P. E. Peterson (Eds.) Earning and learning: How schools matter (pp. 131-168). Washington, DC: Brookings.

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.). (1997). What if there were no significance tests? Hillsdale, NJ: Erlbaum.

Harre, R. (1981). Great scientific experiments. Oxford, England: Phaidon Press.

Harris, M. J., & Rosenthal, R. (1985). Mediation of inrerpersonal expectancy effects: 31 meta-analyses. Psychological Bulletin, 97, 363-386.

Harris, R. J. (1997). Significance tests have their place. Psychological Science, 8, 8-11.

Harrop, J. W., & Velicer, W. F. (1990a). Computer programs for interrupted time series analysis: I. A qualitative evaluation. Multivariate Behavioral Research, 25, 219-231.

Harrop, J. W., & Velicer, W. F. (1990b). Computer programs for interrupted time series analysis: II. A quantitative evaluation. Multivariate Behavioral Research, 25, 233-248.

Hart, H. L. A., & Honore, T. (1985). Causation in the law. Oxford, England: Clarendon Press.

Hartman, R. S. (1991). A Monte Carlo analysis of alternative estimators in models involving selectivity. Journal of Business and Economic Statistics, 9, 41-49.

Hartmann, D, P., & Hall, R. V. (1976). The changing criterion design. Journal of Applied Behavior Analysis, 9, 527-532.

Hartmann, G. W. (1936). A field experiment on the comparative effectiveness of "emotional" and "rational" political leaflets in determining election results. Journal of Abnormal and Social Psychology, 31, 99-114.

Harvey, A. (1990). The econometric analysis of time series (2nd ed.). Cambridge, MA: MIT Press.

Harwell, M. (1997). An empirical study of Hedges's homogeneity test. Psychological Methods, 2, 219-231.

Hathaway, S. R., & McKinley, J. C. (1989). MMPI-2: Manual for Administration and Scaring. Minneapolis: University of Minnesota Hess.

Hauk, W. W., & Anderson, S. (1986). A proposal for interpreting and reporting negative studies. Statistics in Medicine, 5, 203-209.

Hausman, J. A., & Wise, D. A. (Eds.). (1985). Social experimentation. Chicago: Uriiversity of Chicago Press.

Havassey, B. (1988). Efficacy of cocaine treatments: A collaborative study (NIDA Grant Number DA05582). San Francisco: University of California.

Haveman, R. H. (1987). Poverty policy and poverty research: The Great Society and the social sciences. Madison: University of Wisconsin Press.

Hayduk, L.A. (1987). Structural equation modeling with LISREL. Baltimore: Johns Hopkins University Press.

Haynes, R. B., Taylor, D. W., & Sackett, D. L. (Eds.). (1979). Compliance in health care. Baltimore: Johns Hopkins University Press.

Heap, J. L (1995). Constructionism in the rhetoric and practice of Fourth Generation Evaluation. Evaluation and Program Planning, 18, 51-61.

Hearst, N., Newman, T., & Hulley, S. (1986). Delayed effects of the military draft on mortality: A randomized natural experiment. New England Journal of Medicine, 314, 620-634.

Heckman, J. J. (1979). Sample selection bias as a specification error. Econometrica, 47, 153-161.

Heckman, J. J. (1992). Randomization and social policy evaluation. In C. F. Manski & I. Garfinkel (Eds.), Evaluating welfare and training programs (pp. 201-230). Cambridge, MA: Harvard University Press.

Heckman, J. J. (1996). Comment. Journal of the American Statistical Association, 91, 459-462.

Heckman, J. J., & Hotz, V. J. (1989a). Choosing among alternative nonexperimental methods for estimating the impact of social programs: The case of manpower training. Journal of the American Statistical Association, 84, 862-874.

Heckman, J. J., & Hotz, V. J. (1989b). Rejoinder. Journal of the American Statistical Association, 84, 878-880.

Heckman; J. J., Hotz, V. J., & Dabos, M. (1987). Do we need experimental data to evaluate the impact of manpower training on earnings? Evaluation Review, 11, 395-427.

Heckman, J· J., lchimura, H., & Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. Review ofEconomic Studies, 64, 605-654.

Heckman, J· J., LaLonde, R. J., & Smith, J. A. (1999). The economics and econometrics of active labor market programs. In A. Ashenfelter & D. Card (Eds.), Handbook of labor ecoriomics (Vol. 3, pp. 1-160). Amsterdam: Elsevier Science.

Heckman, J. J., & Robb, R. (1985). Alternative methods for evaluating the impact of interventions. InJ. J. Heckman & B. Singer (Eds.), Longitudinal analysis of labor market data (pp. 156-245), Cambridge, England: Cambridge University Press.

Heckman, J. J., & Robb, R. (1986a). Alternative methods for solving the problem of selection bias in evaluating the impact of treatments on outcomes. In H. Wainer (Ed.), Drawing inferences from self-selected samples (pp. 63--107). New York: Springer-Verlag.

Heckman, J. J., & Robb, R. (1986b). Postscript: A rejoinder to Tukey. In H. Wainer (Ed.), Drawing inferences from self-selected samples (pp. 111-113). New York: Springer-Verlag.

Heckman, J. J., & Roselius, R. L. (1994, August). Evaluating the impact of training on the earnings and labor force status of young women: Better data help a lot. (Available from the Department of Economics, University of Chicago)

Heckman, J· J., & Roselius, R. L. (1995, August). Non-experimental evaluation of job training programs for young men. (Available from the Department of Economics, University of Chisago)

Heckman, J·J., & Todd, P. E. (1996, December). Assessing the performance of alternative estimators of program impacts: A study of adult men and women in JTPA. (Available from the Department of Economics, University of Chicago)

Hedeker, D., & Gibbons, R. D. (1994j. A random-effects ordinal regression model for multilevel data. Biometrics, 50, 933-944.

Hedges, L. V. (1984). Estimation of effect size under nonrandom sampling: The effects of censoring studies yielding statistidilly insignificant mean differences. Journal of Educational Statistics, 9, 61-85.

Hedges, L. V. (1992). Modeling publication selection effects in meta-analysis. Statistical Science, 7, 246-255.

Hedges, L. V. (1994). Fixed effects models. In H. Cooper & L. V. Hedges (Eds.), The handbook of research synthesis (pp. 285-299). New York: Russell Sage Foundation.

Hedges, L. V. (1997a). The promise of replication in labour economics. Labour Economics, 4, 111-114.

Hedges, L. V. (1997b). The role of construct validity in causal generalization: The concept of total causal inference error. In V. R. McKim & S. P. Turner (Eds.), Causality in crisis? Statistical methods and the search for causal knowledge in the social sciences {pp. 325-341). Notre Dame, IN: University of Notre Dame Press.

Hedges, L. V., & Olkin, I. (1985). Statistical methods for meta-analysis. Orlando, FL: Academic Press.

Hedges, L. V., & Olkin, I. (in press). Statistical methods for meta-analysis in the medical and social sciences. Orlando, FL: Academic Press.

Hedges, L. V., & Vevea, J. L. (1996). Estimating effect size under publication bias: Small sample properties and robustness of a random effects selection model. Journal of Educational and Behavioral Statistics, 21, 299-332.

Hedges, L. V., & Vevea, J, L. (1998). Fixed- and random-effects models in meta-analysis. Psychologiwl Methods, 4, 486-504.

Heider, F. (1944). Social perception and phenomenal causality. Psychological Review, 51, 358-374.

Heinsman, D. T., & Shadish, W. R. (1996). Assignment methods in experimentation: When do nonrandomized experiments approximate the answers from randomized experiments? Psychologiwl Methods, 1, 154-169.

Heitjan, D. F. (1999). Causal inference in a clinical trial: A comparative example. Controlled Clinical Trials, 20, 309-318.

Hennigan, K. M., Del Rosario, M. L., Heath, L., Cook, T. D., Wharton, J.D., & Calder, B. J. (1982). Impact of the introduction of television on crime in the United States: Empirical findings and theoretical implications. Journal of Personality and Social Psychology, 55, 239-247.

Henry, G. T., & McMillan, J. H. (1993). Performance data: Three comparison methods. Evaluation Review, 17, 643-652.

Herbst, A., Ulfelder, H., & Poskanzer, D. (1971). Adenocarcinoma of the vagina: Association of maternal stilbestrol therapy with tumor appearance in young women. New England Journal of Medicine, 284, 878-881.

Herrnstein, R. J., & Murray, C. (1994). The bell curve. New York: The Free Press.

Hill, C. E., O'Grady, K. E., & Elkin, I. (1992). Applying the Collaborative Study Psychotherapy Rating Scale to rate therapist adherence in cognitive-behavior therapy, interpersonal therapy, and clinical management. Journal of Consulting and Oinical Psychology, 60, 73-79.

Hill, J. L., Rubin, D. B., & Thomas, N. (2000). The design of the New York School Choice Scholarship program evaluation. In L. Bickman (Ed.), Validity and social experimentation: Donald Campbell's legacy (Vol. 1, pp. 155-180). Thousand Oaks, CA: Sage.

Hillis, A., Rajah, M. H., Baisden, C. E., Villamaria, F. J., Ashley, P., & Cummings, C (1998). Three years of experience with prospective randomized effectiveness studies.Controlled Cliniwl Trials, 19, 419-426.

Hillis, J. W., & Wortman, C. B. (1976). Some determinants of public acceptance of randomized control group experimental designs. Sociometry, 39, 91-96.

Hintze, J. L. (1996). PASS User's Guide: PASS 6.0 Power Analysis and Sample Size for Windows. (Available from Number Cruncher Statistical Systems, 329 North 1OOO East, Kaysville, Utah 84037)

Hogg, R. V., & Tanis, E. A. (1988). Probability and statistical inference (3rd ed.). New York: Macmillan.

Hohmann, A. A., & Parron, D. L. (1996). How the new NIH guidelines on inclusion of women and minorities apply: Efficacy trials, effectiveness trials, and validity. Journal of Consulting and Clinical Psychology, 64, 851-855.

Holder, H. D., & Wagenaar, A. C. (1994). Mandated server training and reduced alcohol-involved traffic crashes: A time series analysis of the Oregon experience. Accident Analysis and Prevention, 26, 89-97.

Holland, P. W. (1986). Statistics and causal inference. Journal of the American Statistical Association, 81, 945-970.

Holland, P. W. (1988). Causal inference, path analysis, and recursive structural equation models. In C. C. Clagg (Ed.), Sociological methodology (pp. 449-493). Washington, DC: American Sociological Association.

Holland, P. W. (1989). Comment: It's very clear. Journal of the American Statistical Association, 84, 875-877.

Holland, P. W. (1994). Probabilistic causation without probability. In P. Humphreys (Ed.), Pattrick Suppes: Scientific philosopher (Vol. 1, pp. 257-292). Dordrecht, Netherlands: Kluwer.

Holland, P. W., & Rubin, D. B. (1983). On Lord's paradox. In H. Wainer & 5. Messick (Eds.), Principles of modern psychological measurment (pp. 3-25). Hillsdale, NJ: Erlbaum.

Holland, P. W., & Rabin, D. B. (1988). Causal inference ill retrospective studies. Evaluation Review, 12. 203-231.

Hollister, R. G., & Hill, J. (1995). Problems in the evaluation of community-wide initiatives.In J.P.Connell, A. C. Kubisch, L. B. Schorr, & C. H. Weiss (Eds.), New approaches to evaluating community initiatives: Concepts, methods, and contexts (pp. 127-172). Washington, DC: Aspen Institute.

Holton, G. (1986). The advancement of science, and its burdens. Cambridge, England: Cambridge University Press.

Hopson, R. K. (Ed.). (2000). How and why language matters in evaluation. San Francisco: Jossey-Bass.

Horn, J. L. (1991). Comments on "Issues in factorial invariance." In L. M. Collins & J. L. Horn (Eds.), Best methods for the analysis of change: Recent advances, unanswered questions, future directions (pp. 114-125). Washington, DC: American Psychological Association.

Horowich, P. (1990). Truth. Worcester, Englapd: Basil Blackwell.

Houts, A., & Gholson, B. (1989). Brownian notions: One historicist philosopher's resistance to psychology of science via three truisms and ecological validity. Social Epistemology, 3, 139-146.

Howard, G. S., Maxwell, 5. E., & Fleming, K. J. (2000). The proof of the pudding: An illustration of the relative strengths of null hypothesis, meta-analysis, and Bayesian analysis. Psychological Methods, 5, 315-332'.

Howard, G. S., Millham, j., Slaten, S., & O'Donnell, L. (1981). The effect of subject response style factors on retrospective measures. Applied Psychological Measurement, 5, 89-100.

Howard, G. S., Ralph, K. M., Gulanick, N. A., Maxwell, S. E., Nance, D. W., & Gerber, S. K. (1979). Internal invalidity in pretest-posttest self-reports and a re-evaluation of retrospective pretests. Applied Psychological Measurement, 3, 1-23.

HowArd, K. I., Cox, W. M., & Saunders, S.M. (1988). Attrition in substance abuse comparative treatment research: The illusion of randomization. In L. S. Onken &J.D. Blaine (Eds.), Psychotherapy and counseling in the treatment of drug abuse (pp. 66-79). Rockville, MD: National Institute on Drug Abuse.

Howard, K. I., Kopta, S.M., Krause, M. S., & Orlinsky, D. E. (1986). The dose-effect relationship in psychotherapy. American Psychologist, 41, 159-164.

Howard, K. I., Krause, M.S., & Orlinsky, D. E. {1986). The attrition dilemma: Toward a new strategy for psychotherapy research. Journal of Consulting and Clinical Psychology, 54, 106-110.

Hox, J. J. (1995). AMOS, EQS, and LISREL for Windows: A comparative review. Structural Equation Modeling, 2, 79-91.

Hrobjartsson, A., Gotzche, P. C., & Gluud, C. (1998). The controlled clinical trial turns 100: Fibieger's trial of serum treatment of diptheria. British Medical Journal, 317, 1243-1245.

Hsiao, C. (1986). Analysis of panel data. New York: Cambridge University Press.

Hsiao, C., Lahiri, K., Lee, L.-F., & Pesaran, M. H. (Eds.). (1999). Analysis ofpanels and limited dependent variable models: In honour of G. S. Maddala. Cambridge, England Cambridge University Press.

Huitema, B. E. (1980). The analysis of covariance and alternatives. New York: Wiley.

Hultsch, D. F., & Hickey, T. (1978). External validity in the study of human development: Theoretical and methodological issues. Human Development, 21, 76-91.

Humphreys, P. (Ed.). (1986a). Causality in the social sciences [Special issue]. Synthese, 68(1).

Humphreys, P. (1989). The chances ofexplanation: Causal explanation in the social, medical, and physical sciences. Princeton, NJ: Princeton University Press.

Hunter, J. E. (1997). Needed: A ban on significance tests. Psychological Science, 8, 3-7.

Hunter, J. E., & Schmidt, F. L. (1990). Methods ofmeta-analysis: Correcting error and bias in research findings. Newbury Park CA: Sage.

Hunter, J. E., & Schmidt, F. L. (1994). Correcting for sources of artificial variation across studies. In H. Cooper & L. V. Hedges (Eds.), The handbook of research synthesis (pp. 323-336). New York: Russell Sage Foundation.

Imbens, G. W., & Rubin, D. B. (1997a). Bayesian inference for causal effects in randomized experiments with non-compliance. Annals of Statistics, 25, 305-327.

Imbens, G. W., & Rubin, D. B. (1997b). Estimating outcome distributions for compliers in instrumental variables models. Review of Economic Studies, 64, 555-574.

Imber, S.D., Pilkonis, P. A., Sotsky, S.M., Elkin, I., Watkins, J. T., Collins,J. F., Shea, M. T., Leber, W. R., & Glass, D. R. (1990). Mode-specific effects among three treatments for depression. Journal of Consulting and Clinical Psychology, 58, 352-359.

Innes, J. M. (1979). Attitudes towards randomized control group experimental designs in the field of community welfare. Psychological Reports, 44, 1207-1213.

International Conference on Harmonization. (1999, May 7). Draft consensus guideline: Choice of control group in clinical trials [On-line]. Available: http://www.ifpma.org/ichl.html, or from ICH Secretariat, do IFPMA, 30 rue de St-Jean, P.O. Box 9, 1211 Geneva 18, Switzerland.

Joannidis, J. P. A., Dixon, D. O., Mcintosh, M., Albert, J. M., Bozzette, S. A., & Schnittman, S.M. (1999). Relationship between event rates and treatment effects in clinical site differences within mnlticenter trials: An example from primary Pneumocystic carinii prophylaxis. Controlled Clinical Trials, 20, 253-266.

Isserman, A., & Rephann, T. (1995). The economic effects of the Appalachian Regional Commission: An empirical assessment of 26 years of regional development planning. Journal of the Ameriam Planning Association, 61, 345-364.

Iyengar, S., & Greenhouse, J. B. (1988). Selection models and the file-drawer problem [with discussion]. Statistical Science, 3, 109-135.

Jacobson, N. S., & Baucom, D. H. (1977). Design and assessment of nonspecific control groups in behavior modification research. Behavior Therapy, 8, 709-719.

Jacobson, N. S., Follette, W. C., & Revenstorf, D. (1984). Psychotherapy outcome research: Methods for reporting variability and evaluating clinical significance. Behavior Therapy, 15, 336-352.

Jacobson, N. S., Schmaling, K. B., & Holtzworth-Munroe, A. (1987). Component analysis of behavioral marital therapy: Two-year follow-up and prediction of relapse. Journal of Marital and Family Therapy, 13, 187-195.

Jadad, A. R., Moore, A., Carroll, D., Jenkinson, C., Reynolds, D. J. M., Gavaghan, D. J., & McQuay, H. J. (1996). Assessing the quality of reports of randomized clinical trials: Is blinding necessary? Controlled Clinirol Trials, 17, 1-12.

James, L. R., & Brett, J. M. (1984). Mediators, moderators, and tests for mediation. Journal of Applied Psychology, 69, 307-321.

Jason, L.A., McCoy, K., Blanco, D., & Zolik, E. S. (1981). Decreasing dog litter: Behavioral consultation to help a community group. In H. E. Freeman & M. A. Solomon (Eds.), Evaluation studies review annual (Vol. 6, pp. 660-674). Thousand Oaks, CA: Sage.

Jennrich, R. I., & Schlueter, M. D. (1986). Unbalanced repeated measures models with structured covariance matrices. Biometrics, 42, 805-820.

Jensen, K. B. (1989). Discourses of interviewing: Validating qualitative research findings through textual analysis. In S. Kvale (Ed.), Issues of validity in qualitative research (pp. 93-108). Lund, Sweden: Studentliteratur.

Johnson, B. T. (1989). DSTAT: Software for the meta-analytic review of research litera-tures. Hillsdale NJ: Erlbaum.

Johnson, B. T. (1993). DSTAT 1.10: Software for the meta-analytic review of research literatures [Upgrade documentation]. Hillsdale NJ: Erlbaum.

Johnson, M., Yazdi, K., & Gelb, B. D. (1993). Attorney advertising and changes in the demand for wills. Journal of Advertising, 22, 35-45.

Jones, B. J., & Meiners, M. R. (1986, August). Nursing home discharges: The results of an incentive reimbursement experiment (Long-Term Care Studies Program Research Report; DHHS Publication No. PHS 86-3399). Rockville, MD: U.S. Department of Health and Human Services, Public Health Service, National Center for Health Services Research and Health Care Technology Assessment.

Jones, E. E. (1985). Major developments in social psychology during the past five decades. In G. Lindzey & E. Aronson (Eds.), Handbook of social psychology (Vol. 1, pp. 47-107). New York: Random House.

Jones, E. E., & Sigall, H. (1971). The bogus pipeline: A new paradigm for measuring affect and attitude. Psychological Bulletin, 76, 349-364.

Jones, J. H. (1981). Bad blood: The Tuskegee syphilis experiment. New York: Free Press.

Jones, K. (1991). The application of time series methods to moderate span longitudinal data. In L. M Collins & J. L. Horn (Eds.), Best methods for the analysis of change: Recent advances, unanswered questions, future directions (pp. 75-87). Washington, DC: American Psychological Association.

Jones, W. T. (1969a). A history of Western philosophy: Vol.1. The classirol mind (2nd ed.). New York: Harcourt, Brace, & World.

Jones, W. T. (1969b). A history of Western philosophy: Vol. 3. Hobbes to Hume (2nd ed.). New York: Harcourt, Brace, & World.

Joreskog, K. G., & Sorbom, D. (1988). LISREL 7: A Guide to the Program and Applications. (Available from SPSS, Inc., 444 N. Michigan Ave., Chicago, IL)

Joreskog, K. G., & Sorbom, D. (1990). Model search with TETRAD II and LISREL. Sociologirol Methods and Research, 19, 93-106.

Joreskog, K. G., & Sorbom, D. (1993). LISREL 8: Structural equation modeling with the SIMPLIS command language. (Available from Scientific Software International, Inc., 1525 East 53rd Street, Suite 906, Chicago IL)

Judd, C. M., & Kenny, D. A. (1981a). Estimating the effects of social interventions. Cambridge, England: Cambridge University Press.

Judd, C. M., & Kenny, D. A. (1981b). Process analysis: Estimating mediation in treatment evaluations. Evaluation Review, 5, 602-619.

Judd, C. M., McClelland, G. H., & Culhane, S. E. (1995). Data analysis: Continuing issues in the everyday analysis of psychological data. Annual Review of Psychology, 46, 433-465.

Judge, G., Hill, C., Griffiths, W., & Lee, T. (1985). The theory and practice of econometrics. New York: Wiley.

Kadane, J. B. (Ed.). (1996). Bayesian methods and ethics in clinical trial design. New York: Wiley.

Kalaian, H. A., & Raudenbush, S. W. (1996). A multivariate mixed linear model for meta- analysis. Psychological Methods, 1, 227-235.

Kalish, L.A., & Begg, C. B. (1985). Treatment allocation methods in clinical trials: A review. Statistics in Medicine, 4, 129-144.

Karlin, S. (1987). Path analysis in genetic epidemiology and alternatives. Journal of Edurotional Statistics, 12, 165-177.

Katz, L. F., Kling, J., & Liebman, J. (1997, November). Moving to opportunity in Boston: Early impacts of a housing mobility program. Unpublished manuscript, Kennedy School of Government, Harvard University.

Kazdin, A. E. (1992). Research design in clinirol psychology (2nd ed.). Boston: Allyn & Bacon.

Kazdin, A. E. (1996). Dropping out of child psychotherapy: Issues for research and implications for practice. Clinical Child Psychology and Psychiatry, 1, 133-156.

Kazdin, A. E., & Bass, D. (1989). Power to detect differences between alternative treatments in comparative psychotherapy outcome research. Journal of Consulting and CJinical Psychology, 57, 138-147.

Kazdin, A. E., & Wilcoxon, L.A. (1976). Systematic desensitization and non-specific treatment effects: A methodological evaluation. Psychological Bulletin, 83, 729-758.

Keller, R. T., & Holland, W. E. (1981). Job change: A naturally occurring field experiment. Human Relations, 134, 1053-1067.

Kelling, G. L., Pate, T., Dieckman, D., & Brown, C. E. (1976). The Kansas City Preventive Patrol Experiment: A summary report. In G. V. Glass (Ed.), Evaluation studies review annual (Vol. 1, pp. 605-657). Beverly Hills, CA: Sage.

Kelly, J. A., Murphy, D. A., Sikkema, K. J., McAuliffe, T. L., Roffman, R. A., Solomon, L. J., Winett, R. A., Kalichman, S.C., & The Community HIV Prevention Research Collaborative. (1997). Randomised, controlled, community-level HIV-prevention intervention for sexual-risk behaviour among homosexual men in US cities. Lancet, 350, 1500-1505.

Kendall, M., & Ord, J. K. (1990). Time series (3rd ed.). London: Arnold.

Kendall, P. C. (1998). Empirically supported psychological therapies. Journal of Consulting and Clinical Psychology, 66, 3-6.

Kenny, D. A. (1979). Correlation and causality. New York: Wiley.

Kenny, D. A., & Harackiewicz, J. M. (1979). Cross-lagged panel correlation: Practice and promise. Journal of Applied Psychology, 64, 372-379.

Kenny, D. A., & Judd, C. M. (1984). Estimating the nonlinear and interactive effects of latent variables. Psychological Bulletin, 96, 201-210.

Keppel, G. (1991). Design and analysis: A researcher's handbook (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Kershaw, D., & Fair, J. (1976). The New Jersey income-maintenance experiment: Vol. 1. Operations, surveys, and administration. New York: Academic Press.

Kershaw, D., & Fair, J. (1977). The New Jersey income-maintenance experiment: Vol. 3. Expenditures, health, and social behavior. New York: Academic Press.

Kiecolt, K. J., & Nathan, L. E. (1990). Secondary analysis of suroey data. Thousand Oaks, CA: Sage.

Kim, J., & Trivedi, P. K. (1994). Econometric time series analysis software: A review. American Statistician, 48, 336-346.

Kirk, J., & Miller, M. L. (1986). Reliability and validity in qualitative research. Thousand Oaks, CA: Sage.

Kirk, R. E. (1982). Experimental design: Procedures for the behavioral sciences (2nd ed.).

Belmont, CA: Brooks/Cole.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. Educational and Psychological Measurement, 56, 746-759.

Kirkhart, K. E. (1995). Seeking multicultural validity: A postcard from the road. Evaluation Practice, 16, 1-12.

Kirsch, I. (1996). Hypnotic enhancement of cognitive-behavioral weight loss treatments: Another meta-reanalysis. Journal of Consulting and Clinical Psychology, 64, 517-519.

Kish, L. (1987). Statistical design for research. New York: Wiley.

Kisker, E. E., & Love, J. M. (1999, December). Leading the way: Characteristics and early experiences of selected Early Head Start programs. Washington, DC: U.S. Department of Health and Human Services, Administration on Children, Youth, and Families.

Kitzman, H., Olds, D. L., Henderson, C. R., Hanks, C., Cole, R., Tatelbaum, R., McConnochie, K. M., Sidora, K., Luckey, D. W., Shaver, D., Engelhardt, K., James, P., & Barnard, K. (1997). Effect of prenatal and infancy home visitation by nurses on pregnancy outcomes, childhood injuries, and repeated childbearing. A randomized controlled trial. Journal of the American Medical Association, 278, 644-652.

Klein, L. R. (1992). Self-concept, enhancement, computer education and remediation: A study of the relationship between a multifaceted intervention program and academic achievement. Dissertation Abstracts International, 53 (05), 1471A. (University Microfilms No. 9227700)

Kleinbaum, D. G., Kupper, L. L., & Morgenstern, H. (1982). Epidemiologic research: Principles and quantitative methods. New York: Van Nostrand Reinhold.

Klesges, R. C., Brown, K., Pascale, R. W., Murphy, M., Williams, E., & Cigrang, J. A. (1988). Factors associated with participation, attrition, and outcome in a smoking cessation program at the workplace. Health Psychology, 7, 575-589.

Klesges, R. C., Haddock, C. K., Lando, H., & Talcott, G. W. (1999). Efficacy of a forced smoking cessation and an adjunctive behavioral treatment on long-term smoking rates. Journal of Consulting and Clinical Psychology, 67, 952-958.

Klesges, R. C., Vasey, M. M., & Glasgow, R E. (1986). A worksite smoking modification competition: Potential for public health impact. American Journal of Public Health, 76, 198-200.

Kline, R. B., Canter, W. A., & Robin, A. (1987). Parameters of teenage alcohol abuse: A path analytic conceptual model. Journal of Consulting and Clinical Psychology, 55, 521-528.

Knatterud, G. L., Rockhold, F. W., George, S. L., Barton, F. B., Davis, C. E., Fairweather, W.R.,Honohan,T.,Mowery,R.,& O'Neill,R.(1998).Guidelines for quality assurance in multicenter trials: A position paper. Controlled Clirzical Trials, 19, 477-493.

Knight, G. P., Fabes, R. A., & Higgins, D. A. (1996). Concerns about drawing causal inferences from meta-analyses: An example in the study of gender differences in aggression. Psychological Bulletin, 119,410-421.

Knorr-Cetina, K.D. (1981). The manufacture of knowledge: An essay on the constructivist and contextual nature of science. Oxford, England: Pergamon.

Koehler, M. J., & Levin, J. R. (1998). Regulated randomization: A potentially sharper analytical tool for the multiple baseline design. Psychological Methods, 3, 206-217.

Koepke, D., & Flay, B. R. (1989). Levels of analysis. In M. T. Braverman (Ed.), Evaluating health promotion programs (pp. 75-87). San Francisco: Jossey-Bass.

Koepsell, T. D., Martin, D: C., Diehr, P. H., Psaty, B. M., Wagner, E. G., Perrin, E. B., & Cheadle, A. (1991). Data analysis and sample size issues in evaluations of community- based health promotion and disease prevention programs: A mixed-model analysis of variance approach. Journal of Clinical Epidemiology, 44, 701-713.

Kollins, S. H., Newland, M. C., & Critchfield, T. S. (1999). Quantitative integration of single-subject studies: Methods and misinterpretations. Behavior Analyst, 22, 149-157.

Kopta, S.M., Howard, K.l., Lowry, J. L., & Beutler, L. E. (1994). Patterns of symptomatic recovery in psychotherapy. Journal of Consulting and Clinical Psychology, 62, 1009-1016.

Korfmacher, J., O'Brien, R., Hiatt, s., & Olds, D. (1999). Differences in program implementation between nurses and paraprofessionals providing home visits during nregnancy and infancy: A randomized trial. American Journal of Public Health, 89, 1847-1851.

Koricheva, J., Larsson, S., & Haukioja, E. (1998). Insect performance on experimentally stressed woody plants: A meta-analysis. Annual Review of Entomology, 43, 195-216.

Kraemer, H. C., Gardner, C., Brooks, J. L., & Yesavage, J. A. (1998). Advantages of excluding underpowered studies in meta-analysis: Inclusionist versus exclusionist viewpoints. Psychological Methods, 3, 23-31.

Kraemer, H. C., & Thiemann, S. (1989). A strategy to use soft data effectively in randomized controlled clinical trials. Journal of Consulting and Clinical Psychology, 57, 148-154.

Krantz, D. H. {1999). The null hypothesis testing controversy in psychology. Journal of the American Statistical Association, 94, 1372-1381.

Kratochwill, T. R., & Levin, J, R. (1992). Single-case research design and analysis: New directions for psychology and education. Hillsdale, NJ: Erlbaum.

Kromrey, J.D., & Foster-Johnson, L. {1996). Determining the efficacy of intervention: The use of effect sizes for data analysis in single-subject research. Journal of Experimental Education, 65, 73-93.

Kruglanski, A. W., & Kroy, M. (1976). Outcome validity in experimental research: A reconceptualization. Journal of Representative Research in Social Psychology, 7, 168-178.

Krull, J. L., & MacKinnon, D.P. (1999). Multilevel mediation modeling in group-based intervention studies. Evaluation Review, 23, 418-444.

Kruse, A. Y., Kjaergard, L. L., Krogsgaard, K., Gluud, C., Mortensen, E. L., Gottschau, A., Bjerg, A., and the INFO Trial Group. {2000): A randomized trial assessing the impact of written information on outpatients' knowledge about and attitude toward randomized clinical trials. Controlled Clinical Trials, 21, 223-240.

Kruskal, W., & Mosteller, F. (1979a). Representative sampling: I. Non-scientific literature. International Statistical Review, 47, 13-24.

Kruskal, W., & Mosteller, F. {1979b). Representative sampling: II. Scientific literature, excluding statistics. International Statistical Review, 47, 111-127.

Kruskal, W., & Mosteller, F. (1979c). Representative sampling: III. The current statistical literature. International Statistical Review, 47, 245-265.

Kuhn, T. S. (1962). The structure of scientific revolutions. Chicago: University of Chicago Pfess.

Kunz, R., & Oxman, D. (1998). The unpredictability paradox: Review of empirical comparisons of randomised and non-randomised clinical trials. British Medical Journal, 317, 1185-1190.

Kvale, S. (Ed.). (1989). Issues of validity in qualitative research. Lund, Sweden: Studentlitteratur.

Lachin, J. M. (1988). Statistical properties of randomizarion. In clinical trials. Controlled Clinical Trials, 9, 289-311.

Lachin, J. M. {2000). Statistical considerations in the intent-to-treat principle. Controlled Clinical Trials, 21; 167-189.

Lachin, J. M., Matts, J.P., & Wei, L. J. (1988). Randomization in clinical trials: Conclusions and recommendations. Statistics in Medicine, 9, 365-374.

Lakatos, I. (1978). The methodology of scientific research programmes. Cambridge, England: Cambridge University Press.

Lakoff, G. (1985). Women, fire, and dangerous things. Chicago: University of Chicago Press.

LaLonde, R. (1986). Evaluating the econometric evaluations of training programs with experimental data. American Economic Review, 76, 604-620.

LaLonde, R., & Maynard, R. (1987). How precise are evaluations of employment and training experiments: Evidence from a field experiment. Evaluation Review, 11, 428--451.

Lam, J. A., Hartwell, S. W., & Jekel, J. F. (1994). "I prayed real hard, so I know I'll get in": Living with randomization. In K. J. Conrad (Ed.), Critically evaluating the role of experiments (pp. 55-66). San Francisco: Jossey-Bass.

Lana, R. C. (1969). Pretest sensitization. In R. Rosenthal & R. L. Rosnow (Eds.), Artifact in behavioral research (pp. 119-141). New York: Academic Press.

Larson, R. C. (1976). What happened to patrol operations in Kansas City? Evaluation, 3, 117-123.

Latane, B., & Darley, J. M. (1970). The unresponsive bystander: Why doesn't he help? New York: Appleton-Century-Crofts.

Latane, B., & Nida, S. (1981). Ten years of research on group size and helping. Psychological Bulletin, 89, 308-324.

Latour, B. (1987). Science in action. Cambridge, MA: Harvard University Press.

Latour, B., & Woolgar, S. {1979). Laboratory life: The social construction of scientific facts. Beverly Hills, CA: Sage.

Lavori, P. W. (1992). Clinical trials in psychiatry: Should protocol deviation censor patient data? Neuropsychopharmacology, 6, 39--48.

Lavori, P. W., Louis, T. A., Bailar, J. C., & Polansky, H. (1986). Designs for experiments: Parallel comparisons of treatment. In J. C. Bailar & F. Mosteller (Eds.), Medical uses of .statistics (pp. 61-82). Waltham, MA: New England Journal of Medicine.

Lazar, I., & Darlington, R. (1982). Lasting effects of early education. Monographs of the Society for Research in Child Development, 47 (2-3, Serial No. 195).

Lazarsfeld, P. E (1947). The mutual effects ofstatistical variables. Unpublished manuscript, Columbia University, Bureau of Applied Social Research.

Lazarsfeld, P. F. (1948). The use of panels in social research. Proceedings of the American Philosophical Society, 92, 405--410.

Leaf, R. C., DiGiuseppe, R., Mass, R., & Alington, D. E. (1993). Statistical methods for analyses of incomplete service records: Concurrent use of longitudinal and crosssectional data. Journal of Consulting and Clinical Psychology, 61, 495-505.

Lee, S., & Hershberger, S. (1990). A simple rule for generating equivalent models in covariance structure modeling. Multivariate Behavioral Research, 25, 313-334.

Lee, Y., Ellenberg, J., Hirtz, D., & Nelson, K. (1991). Analysis of clinical trials by treatment actually received: Is it really an option? Statistics in Medicine, 10, 1595-1605.

Lehman, D. R., Lempert, R. O., & Nisbett, R. E. (1988). The effects of graduate training on reasoning: Formal discipline and thinking about everyday-life events. American Psychologist, 43, 431-442.

LeLorier, J., Gregoire, G., Benhaddad, A., Lapierre, J., & Derderian, F. (1997). Discrepancies between meta-analyses and subsequent large randomized, controlled trials. New England Journal of Medicine, 337, 536-542.

Leviton, I. C., & Cook, T. D. (1983). Evaluation findings in education and social work textbooks. Evaluation Review, 7, 497-518.

Leviton, L. C., Finnegan, J. R., Zapka, J. G., Meischke, H., Estabrook, B., Gilliland, J., Linares, A., Weitzman, E. R., Raczynski, J., & Stone, E. (1999). Formative research methods to understand patient and provider responses to heart attack symptoms. Evaluation and Program Planning, 22, 385-397. '

Levy, A. S., Mathews, O., Stephenson, M., Tenney, J. E., & Schucker, R. E. (1985). The.impact of a nutrition information program on food purchases. Journal of Public Policy and Marketing, 4, 1-13.

Lewin, K. (1935). A dynamic theory of personality: Selected papers. New York: McGraw-Hill.

Lewis, D. (1973). Causation. Journal of Philosophy, 70,556-567.

Lewontin, R. (1997, January 9). Billions and billions of demons. New York Review of Books, 64(1), 28-32.

Li, Z., & Begg, C. B. (1994). Random effects models for combining results from controlled and uncontrolled studies in a meta-analysis. Journal of the American Statistical Association, 89, 1523-1527.

Lichstein, K. L. (1988). Clinical relaxation strategies. New York: Wiley.

Lichstein, K. L., Riedel, B. W., & Grieve, R. (1994). Fair tests of clinical trials: A treatment· implementation model. Advances in Behavior Research and Therapy, 16, 1-29.

Lichtenstein, E. L., Glasgow, R. E., & Abrams, D. B. (1986), Social support in smoking cessation: In search of effective interventions. Behavior Therapy, 17, 607-619.

Lichtenstein, L. M. (1993). Allergy and the immune system. Scientific American, 269, 117-124.

Lieberman, S. (1956). The effects of changes in roles on the attitudes of role occupants. Human Relations, 9, 385-402.

Lieberson, S. (1985). Making it count: The improvement of social research and theory. Berkeley: University of California Press.

Liebman, B. (1996). Vitamin E and fat: Anatomy of a flip-flop. Nutrition Action Newsletter, 23, 10-11.

Light, R. J., Singer, J. D., & Willett, J. B. (1990). By design: Planning research in higher education. Cambridge, MA: Harvard University Press.

Light, R. J., Singer, J.D., & Willett, J. B. (1994). The visual presentation and interpretation of meta-analysis. In H. Cooper & L. V. Hedges (Eds.), The handbook of research synthesis (pp. 439-453). New York: Russell Sage Foumlation.

Lincoln, Y. S. (1990). Campbell's retrospective and a constructivist's perspective. Haroard Educational Review, 60, 501-504.

Lincoln, Y. S., & Guba, E. G. (1985). Naturalistic inquiry. Newbury Park, CA: Sage.

Lind, E. A. (1985). Randomized experiments in the Federal courts. In R. F. Boruch & W. Wothke (Eds.), Randomization and field experimentation (pp. 73-80). San Francisco: Jessey-Bass.

Lind, J. (1753). A treatise of the scurvy. Of three parts containing an inquiry into the nature, causes and cure of that disease. Edinburgh: Sands, Murray, & Cochran.

Lipsey, M. W. (1990). Design sensitivity: Statistical power (or experimental research. Thousand Oaks, CA: Sage.

Lipsey, M. W. (1992). Juvenile delinquency treatment: A meta-analytic inquiry into the variability of effects. InT. D. Cook, H. M. Cooper, D. S. Cordray, H. Hartmann, L. V. Hedges, R. J. Light, T. A. Louis, & F. Mosteller (Eds.), Meta-analysis for explanation: A casebook (pp. 83-127). New York: Russell Sage Foundation.

Lipsey, M. W., Cordray, D. S., & Berger, D. E. (1981). Evaluation of a juvenile diversion program. Evaluation Review, 5, 283-306.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. American Psychologist, 48, 1181-1209.

Lipsey, M. W., & Wilson, D. B. (2000). Practical meta-analysis. Newbury Park, CA: Sage. Little, R. J. (1985). A note about modelsfor selectivity bias. Econometrica, 53, 1469-1474. Little, R. J. (1995). Modeling the drop-out mechanism in repeated-measures studies. Journal of the American Statistical Association, 90, 1112-1121.

Little, R. J., & Rubin, D. B. (1987). Statistical analysis with missing data. New York: Wiley.

Little, R. J., & Rubin, D. B. (1999). Comment. Journal ofthe American Statistical Association, 94, 1130-1132.

Little, R. J., & Schenker, N. (1995). Missing data. In G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), Handbook of statistical modeling for the social and behavioral sciences (pp. 39-75). New York: Plenum Press.

Little, R. J., & Yau, L. (1996). Intent-to-treat analysis for longitudinal studies with dropouts. Biometrics, 52, 1324-1333.

Little, R. J., & Yau, L. (1998). Statistical techniques for analyzing data from prevention trials: .Treatment of no-shows using Rubin's causal model. Psychological Methods, 3, 147-159.

Locke, E. A. (1986). Generalizing from laboratory to field settings. Lexington, MA: Lexington Books.

Locke, H. J., & Wallace, K. M. (1959). Short-term marital adjustment and prediction tests: Their reliability and validity. Journal of Marriage and Family Living, 21, 251-255.

Locke, J. (1975). An essay concerning human understanding. Oxford, England: Clarendon Press. (Original work published in 1690)

Lockhart, D. C. (Ed.). (1984). Making effective use of mailed questionnaires. San Francisco: Jossey-Bass.

Loehlin, J. C. (1992, January). Latent variable models: An introduction to factor, path, and structural analysis (2nd ed.). Hillsdale, NJ: Erlbaum.

Lohr, B. W. (1972). An historical view of the research on the factors related to the utilization of health services. Unpublished mauuscript. Rockville, MD: Bureau for Health Services Research and Evaluation, Social and Economic Analysis Division.

Looney, M. A., Feltz, C. J., & Van Vleet, C. N. (1994). The reporting and analysis of research findings for within-subject designs: Methodological issues for meta-analysis. Research Quarterly for Exercise & Sport, 65, 363-366.

Lord, F. M. (1963). Elementary models for measuring change. In C. W. Harris (Ed.), Problems in measuring change (pp. 21-38). Madison: University of Wisconsin Press.

Lord, F. M. (1967). A paradox in the interpretation of group comparisons. Psychological Bulletin, 68, 304-305.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA·. Addison-Wesley.

Louis, T. A., Fineberg, H. V., & Mosteller, F. (1985). Findings for public health from meta- analyses. Annual Review of Public Health, 6, 1-20.

Louis, T. A., & Zelterman, D. (1994). Bayesian approaches to research synthesis. In H. Cooper & L. V. Hedges (Eds.), The handbook of research synthesis (pp. 411-422). New York: Russell Sage Foundation.

Ludwig, J., Duncan, G.J., & Hirschfield, P. (1998, September). Urban poverty and juvenile crime: Evidence from a randomized housing-mobility experiment. (Available, from author., Ludwig, Georgetown Public Policy Institute, 3600 N Street Nw, Suite 200, Washington, DC 20007)

Lufr, H. S. (1990). The applicability of the regression discontinuity design in health services research. In L. Sechrest, E. Perrin, & J. Bunker (Eds.), Research methodology: Strengthening causal interpretations of nonexperimental data (pp. 141-143). Rockville, MD: Public Health Service, Agency for Health Care Policy and Research.

Lund, E. (1989). The validity of different control groups in a case-control study: Oral contraceptive use and breast cancer in young women. Journal of Clinical Epidemiology, 42, 987-993.

MacCallum, R. C. (1986). Specification searches in covariance structure modeling. Psychological Bulletin, 100, 107-120.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. Psychological Methods, 1, 130-149.

MacCallum, R. C., Wegener, D. T., Uchino, B. N., & Fabrigar, L. R. (1993). The problem of equivalent models in applications of covariance structure analysis. Psychological Bulletin, 114, 185-199.

MacCready, R. A. (1974). Admissions of phenylketonuric patients to residential institutions before and after screening programs of the newborn infant. Journal of Pediatrics, 85, 383-385.

Macintyre, A. (1981). After virtue. Notre Dame, IN: University of Notre Dame Press.

MacKenzie, A., Funderburk, F. R., Allen, R. P., & Stefan, R. L. (1987). The characteristics of alcoholics frequently lost to follow-up. Journal of Studies on Alcohol, 48, 119-123.

Mackie, J. L. (1974). The cement of the universe: A study of causation. Oxford, England: Oxford University Press.

MacKinnon, D.P., Johnson, C. A., Pentz, M.A., Dwyer, J. H., Hansen, W. B., Flay, B. R., & Wang, E.Y.-I. (1991).Mediating mechanisms in a school-based drug prevention program: First-year effects of the Midwestern Prevention Project. Health Psychology, 10, 164-172.

MacKinnon, D.P., Warsi, G., & Dwyer, J. H. (1995). A simulation study of mediated effect measures. Multivariate Behavioral Research, 30, 41-62.

Madden, E. H., & Humber, J. (1971). Nomological necessity and C. J. Ducasse. Ratio, 13, 119-138.

Madaus, G. F., & Greaney, V. (1985). The Irish Experience in competency testing: Implications for American education. American Journal of Education, 93, 268-294.

Magidson, J. (1977). Toward a causal model approach for adjusting for preexisting differences in the nonequivalent control group situation. Evaluation Quarterly, 1, 399-420.

Magidson, J. (1978). Reply to Bentler and Woodward: The .05 significance level is not all-powerful. Evaluation Quarterly, 2, 511-520.

Magidson, J. (2000). On models used to adjust for preexisting differences. In L. Bickman (Ed.), Research design: Donald Campbell's legacy (Vol. 2, pp. 181-194). Thousand Oaks, CA: Sage.

Magnusson, D. (2000). The individual as the organizing principle in psychological inquiry: A holistic approach. In L. R. Bergman, R. B. Cairns, L. Nilsson, & L. Nystedt (Eds.), Developmental science and the holistic approach (pp. 33--47). Mahwah, NJ: Erlbaum.

Makuch, R., & Simon, R. (1978). Sample size requirements for evaluating a conservative therapy. Cancer Treatment Reports, 62, 1037-1040.

Mallams, J. H., Godley, M.D., Hall, G. M., & Meyers, R. J. (1982). A social systems approach to resocializing alcoholics in the community. Journal of Studies on Alcohol, 43, 1115-1123.

Maltz, M.D., Gordon, A. C., McDowall, D., & McCleary, R. (1980). An artifact in pretest-posttest designs: How it can mistakenly make delinquency programs look effective. Evaluation Review, 4, 225-240.

Mann, C. (1994). Can meta-analysis make policy? Science, 266, 960-962.

Mann, T. (1994). Informed consent for psychological research: Do subjects comprehend consent forms and understand their legal rights? Psychological Science, 5,140-143.

Manski, C. F. (1990). Nonparametric bounds on treatment effects. American Economic Review Papers and Proceedings, 80, 319-323.

Manski, C. F., & Garfinkel, I. (1992). Introduction. In C. F. Manski & I. Garfinkel (Eds.), Evaluating welfare and training programs (pp. 1-22). Cambridge, MA: Harvard University Press.

Manski, C. F., & Nagin, D. S. (1998). Bounding disagreements about treatment effects: A case study of sentencing and recidivism. In A. Raftery (Ed.), Sociological methodology (pp. 99-137). Cambridge, MA: Blackwell.

Manski, C. F., Sandefur, G. D., McLanahan, S., & Powers, D. (1992). Alternative estimates of the effect of family structure during adolescence on high school graduation. Journal of the American Statistical Association, 87, 25-37.

Marascuilo, L.A., & Busk, P. L. (1988). Combining statistics for multiple-baseline AB and replicated ABAB designs across subjects. Behavioral Assessment, 10, 1-28.

Marcanronio. R. J. (1998). ESTIMATE: Statistical software to estimate the impact of missing data Computer software]. Lake in the Hills, IL: Statistical Research Associates.

Marcus, S.M. (1997a). Assessing non-consent bias with parallel randomized and nonrandomized clinical trials. Journal ofClinical Epidemiology, SO, 823-828.

Marcus, S.M. (1997b). Using omitted variable bias to assess uncertainty in the estimation of an AIDS education treatment effect. Journal of Educational and Behavioral Statistics, 22, 193-201.

Marcus, S.M. (2001). A sensitivity analysis for subverting randomization in controlled trials. Statistics in Medicine, 20, 545-555.

Margraf, J., Ehlers, A., Roth, W. T., Clark, D. B., Sheikh, J., Agras, W. S., & Taylor, C. B. (1991). How "blind" are double-blind studies? Journal of Consulting and Clinical Psychology, 59, 184-187.

Marin, G., Marin, B. V., Perez-Stable, E. J., Sabogal, F., & Ostero-Sabogal, R. (1990). Changes in information as a function of a culturally appropriate smoking cessation community intervention for Hispanics. American journal of Community Psychology, 18, 847-864.

Mark, M. M. (1986). Validity typologies and the logic and practice of quasi-experimentaton. In W. M. K. Trochim {Ed.), Advances in quasi-experimental design and analysis (pp. 47-66). San Francisco: Jossey-Bass.

Mark, M. M. (2000). Realism, validity, and the experimenting society. In L. Bickman (Ed.), Validity and social experimentation: Donald Campbell's legacy (Vol. 1, pp. 141-166). Thousand Oaks, CA: Sage.

Mark, M. M., & Mellor, S. (1991). Effect of self-relevance of an event on hindsight bias: The foreseeability of a layoff. Journal of Applied Psychology, 76, 569-577.

Marquis, D. (1983). Leaving tberapy to chance. Hastings Center Report, 13, 40-47.

Marriott, F. H. C. (1990). A dictionary of statistical terms (5th ed.). Essex, England: Longman Scientific and Technical.

Marschak, J. (1953). Economic measurements for policy and prediction. In W. C. Hood & T. C. Koopmans (Ed.), Studies in econometric method (Cowles Commission Monograph No. 13). New York: Wiley.

Marsh, H. W. (1998). Simulation study of non-equivalent group-matching and regression-discontinuity designs: Evaluation of gifted and talented programs. Journal of Experimental Education, 66, 163-192.

Marshall, E. (1989). Quick release of AIDS drugs. Science, 245, 345, 347.

Martin, D. C., Diehr, P., Perrin, E. B., & Koepsell, T. D. (1993). The effect of matching on the power of randomized community intervention studies. Statistics in Medicine, 12, 329-338.

Martin, S. E., Annan, S., & Forst, B. (1993). The special deterrent effects of a jail sanction on first-time drunk drivers: A quasi-experimental study. Accident Analysis and Prevention, 25, 561-568.

Marx, J. L. (1989). Drug availability is an issue for cancer patients, too. Science, 245, 346-347.

Mase, B. F. (1971). Changes in self-actualization as a result of two types of residential group experience. Unpublished doctoral dissertation, Northwestern University, Evanston, IL.

Mastroianni, A. C., Faden, R., & Federman, D. (Eds.). (1994). Women and health research (Vols. 1-2). Washington, DC: National Academy Press.

Matt, G. E. (1989). Decision rules for selecting effect sizes in meta-analysis: A review and reanalysis of psychotherapy outcome studies. Psychological Bulletin, 105, 106-115.

Matt, G. E., & Cook, T. D. (1994). Threats to the validity of research syntheses. In H. Cooper & L. V. Hedges (Eds.), The handbook of research synthesis (pp. 503-520). New York: Russell Sage Foundation.

Matt, G. E., Cook, T. D., & Shadish, W. R. (2000). Generalizing about causal inferences. Manuscript in preparation.

Mauro, R. (1990). Understanding L.O.V.E. (Left Out Variables Error): A method for examining the effects of omitted variables. Psychological Bulletin, 108, 314-329.

Maxwell, J. A (1990). Response to "Campbell's retrospective and a constructivist's perspective." Harvard Educational Review, 60, 504-508.

Maxwell, J. A. (1992}. Understanding and validity in qualitative research. Haroard Educatiqnal Review, 62: 279-300.

Maxwell, J. A., Bashook, P. G., & Sandlow, L. J. (1986). Combining ethnographic and experimental methods in educational evaluation: A case study. In D. M. Fetterman & M.A. Pittman (Eds.), Educational evaluation: Ethnography in theory, practice, and politics (pp.121-143). Newbury Park, CA: Sage.

Maxwell, J. A., & Lincoln, Y. S. (1990). Methodology and epistemology: A dialogue. Harvard Educational Review, 60, 497-512.

Maxwell, S. E. (1993). Covariate imbalance and conditional size: Dependence on modelbased adjustments. Statistics in Medicine, 12, 101-109.

Maxwell, S. E. (1994). Optimal allocation of assessment time in randomized pretest- posttest designs. Psychological Bulletin, 115, 142-152.

Maxwell, S. E. (1998), Longitudinal designs in randomized group comparisons: When will intermediate observations increase statistical power? Psychological Methods, 3, 275-290.

Maxwell, S. E., Cole, D. A., Arvey, R. D., & Salas, E. (1991). A comparison of methods for increasing power in randomized between-subjects designs. Psychological Bulletin, 110, 328-337.

Maxwell, S. E., & Delaney, H. D. (1990). Designing experiments and analyzing data: A model comparison approach. Pacific Grove, CA: Brooks/Cole.

Maxwell, S. E., Delaney, H. D., & Dill, C. A. (1984). Another look at ANCOVA versus blocking. PsycholOgical Bulletin, 95, 136-147.

Mayer, L. S. (1986). Statistical inferences for cross-lagged panel models without the assumption of normal errors. Social Science Research, 15, 28-42.

McAweeney, M. J., & klockars, A. J. (1998). Maximizing power in skewed distributions: Analysis and assignment. Psychological Methods, 3, 117-122.

McCall, W. A. (1923). How to experiment in education. New York: MacMillan.

McCardel, J. B. (1972). Interpersonal effects of structured and unstructured human relations groups. Dissertation Abstracts International, 33, 4518-4519. (University Microfilms No. 73-5828)

McClannahan, L. E., McGee, G. G., MacDuff, G. S., & Krantz, P. J. (1990). Assessing and improving child care: A personal appearance index for children with autism. Journal of Applied Behavior Analysis, 23, 469-482.

McCleary, R. D. (2000). The evolution of the time series experiment. In L. Bickman (Ed.), Research design: Donald Campbell's legacy (Vol. 2, pp. 215-234). Thousand Oaks, CA: Sage.

McCleary, R. D., & Welsh, W. N. {1992). Philosophical and statistical foundations of time-series experiments. InT. R. Kratochwill & J. R. Levin (Eds.), Single-case research design and analysis (pp. 41-91). Hillsdale, NJ: Erlbaum.

McClelland, G. H. (1997). Optimal design in psychological research. Psychological Methods, 2, 3-19.

McClelland, G. H. (2000). Increasing statistical power without increasing sample size. American Psychologist, 55, 963-964.

McClelland, G. H., & Judd, C. M. (1993). Statistical difficulties of detecting interactions and moderator effects. Psychological Bulletin, 114, 376-390.

McCord, J. (1978). A thirty-year followup of treatment effects. American Psychologist, 33, 284-289.

McCord, W., & McCord, J. {1959). Origins of crime. New York: Columbia University Press.

McCoy, H. V., & Nurco, D. M. (1991). Locating subjects by traditional techniques. In D. M. Nurco (Ed.), Follow-up fieldwork: AIDS outreach and IV drng abuse (DHHS Publication No. ADM 91-1736, pp. 31-73). Rockville, MD: National Institute on Drug Abuse.

McCullough, B. D., & Wilson, B. (1999). On the accuracy of statistical procedure in Microsoft Excel 97. Computational Statistics and Data Analysis, 31, 27-37.

McDonald, R. P. (1994). The bilevel reticular action model for path analysis with latent variables. Sociological Methods and Research, 22, 399-413.

McDowall, D., McCleary, R., Meidinger, E. E., & Hay, R. A. (1980). Interrupted time series analysis. Newbury Park CA: Sage.

McFadden, E. (1998). Management of data in clinical trials. New York: Wiley.

McGuire, W. J. (1984). A contextualist theory of knowledge: Its implications for innovation and return in psychological research. In L. Berkowitz (Ed.), Advances in experimental social psychology (pp. 1-47). New York: Academic Press.

McGuire, W. J. (1997). Creative hypothesis generating in psychology: Some useful heuristics. Annual Review of Psychology, 48, 1-30.

McKay, H., Sinisterra, L., McKay, A., Gomez, H., & Lloreda, P. (1978). Improving cognitive ability in chronically deprived children. Science, 200, 270--278.

McKillip, J. (1992). Research without control groups: A control construct design. In F. B. Bryant, J. Edwards, R. S. Tindale, E. J. Posavac, L. Heath, & E. Henderson (Eds.), Methodological issues in applied psychology (pp. 159-175). New York: Plenum.

McKillip, J., & Baldwin, K. (1990). Evaluation of an STD education media campaign: A control construcr design. Evaluation Review, 14, 331-346.

McLeod, R. S., Taylor, D. W., Cohen, A., & Cullen, J. B. (1986, March 29). Single patient randomized clinical trial: Its use in determining optimal treatment for patient with inflammation of a Kock continent ileostomy reservoir. Lancet, 1, 726-728.

McNees, P., Gilliam, S. W., Schnelle, J: F., & Risley, T. (1979). Controlling employee theft through time and product identification. Journal of Organizational Behavior Management, 2, 113-119.

McSweeny, A. J. (1978). The effects of response cost on the behavior of a million persons: Charging for directory assistance in Cincinnati. Journal of Applied Behavioral Analysis, 11, 47-51.

Mead, R. (1988). The design of experiments: Statistical principles for practical application. Cambridge, England: Cambridge University Press.

Medin, D. L. (1989). Concepts and conceptual structures. American Psychologist, 44, 1469-1481.

Meehl, P. E. (1967). Theory testing in psychology and physics: A methodological paradox. Philosophy of Science, 34, 103-115.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. Journal of Consulting and Clinical Psychology, 46, 806-834.

Meier, P. (1972). The biggest public health experiment ever: The 1954 field trial of the Salk poliomyelitis vaccine. In J. M. Tanur, F. Mosteller, W. H. Kruskal, R. F. Liuk, R. S. Pieters, & G. R. Rising (Eds.), Statistics: A guide to the unknown (pp. 120-129). San Francisco: Holden-Day.

Meinert, C. L., Gilpin, A. K, Unalp, A., & Dawson, C. (2000). Gender representation in trials. Controlled Clinical Trials, 21, 462-475.

Menard, S. (1991). Longitudinal research. Thousand Oaks, CA: Sage.

Mennicke, S. A., Lent, R. W., & Burgoyne, K. L. (1988). Premature termination from university counseling centers: A review. Journal of Counseling and Development, 66, 458-464.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement (3rd ed.), (pp. 13-103). New York: Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. American Psychologist, 50, 741-749.

Meyer, B. D. (1995). Natural and quasi-experiments in economics. Journal of Business and Economic Statistics, 13, 151-161.

Meyer, D. L. (1991). Misinterpretation of interaction effects: A reply to Rosnow and Rosenthal. Psychological Bulletin, 110, 571-573.

Meyer, T. J., & Mark, M. M. (1995). Effects of psychosocial interventions with adult cancer patients: A meta-analysis of randomized experiments. Health Psychology, 14, 101-108.

Miettineu, 0. S. (1985). The "case-control" study: Valid selection of subjects. Journal of Chronic Diseases, 38, 543-548.

Mike, V. (1989). Philosophers assess randomized clinical trials: The need for dialogue. Controlled Clinical Trials, 10, 244-253.

Mike, V. (1990). Suspended judgment: Ethics, evidence, and uncertainty. Controlled Clinical Trials, 11, 153-156.

Miles, M B., & Huberman, A.M. (1984). Qualitative data analysis: A sourcebook of new methods. Newbury Park, CA: Sage.

Miller, N., Pedersen, W. C., & Pollock, V. E. (2000). Discriminative validity. In L. Bickman (Ed.), Validity and social experimentation: Donald Campbell's legacy (Vol. 1, pp. 65-99). Thousand Oaks, CA: Sage.

Miller, T. Q., Turner, C. W., Tindale, R. S., Posavac, E. J., & Dugoni, B. L. (1991). Reasons for the trend toward null findings in research on Type A behavior. Psychological Bulletin, 110, 469-485.

Millsap, M. A., Goodson, B., Chase, A., & Gamse, B. (1997, December). Evaluation of «Spreading the Comer School Development Program and Philosophy." Report submitted to the Rockefeller Foundation, 420 Fifth Avenue, New York, NY 10018 by Abt Associates Inc., 55 Wheeler Street, Cambridge MA 02138.

Minton, J. H. (1975). The impact of "Sesame Street" on reading readiness of kindergarten children. Sociology of Education, 48, 141-151.

Mishler, E. G. (1990). Validation in inquiry-guided research: The role of exemplars in narrative studies. Harvard Educational Review, 60, 415-442.

Mitroff, I. I., & Fitzgerald, I. (1977). On the psychology of the Apollo moon scientists: A chapter in the psychology of science. Human Relations, 30, 657-674.

Moberg, D.P., Piper, D. L., Wu, J., & Serlin, R. C. (1993). When total randomization is impossible: Nested random assignment. Evaluation Review, 17, 271-291.

Moerbeek, M., van Breukelen, G. J.P., & Berger, M.P. F. (2000). Design issues for experiments in multilevel populations. Journal of Educational and Behavioral Statistics, 25, 271-284.

Moffitt, R. A. (1989). Comment. Journal of the American Statistical Association, 84, 877-878.

Moffitt, R. A. (1991). Program evaluation with nonexperimental data. Evaluation Review, 15, 291-314.

Moffitt, R. A. (1996). Comment. Journal of the American Statistical Association, 91, 462-465.

Moher, D., Jadad, A. R., Nichol, G., Penman, M., Tugwell, P., & Walsh, S. (1995). Assessing the quality of randomized controlled trials: An annotated bibliography of scales and checklists. Controlled Clinical Trials, 16, 62-73.

Moher, D., & Olkin, I. (1995). Meta-analysis of randomized controlled trials: A concern for standards. Journal of the American Medical Association, 274, 1962-1963.

Mohr, L. B. (1988).Impact analysis for program evaluation. Chicago: Dorsey Press.

Mohr, L. B. (1995). Impact analysis for program evaluation (2nd ed.). Thousand Oaks,

CA: Sage.

Moos, R. H. (1997). Evaluating treatment environments: The quality of psychiatric and substance abuse programs (2nd ed.). New Brunswick, NJ: Transaction.

Morales, A. (2000, November). Investigating rules and principles for combining qualitative and quantitative data. Paper presented at the annual conference of the American Evaluation Association, Honolulu, Hawaii.

Morawski, J. G. (1988). The rise of experimentation in American pSychology. New Haven, CT: Yale University Press.

Mosteller, F. (1990). Improving research methodology: An overview. In L. Sechrest, E. Perrin, & J. Bunker (Eds.), Research methodology: Strengthening causal interpretations of nonexperimental data (pp. 221-230). Rockville, MD: U.S. Public Health Service, Agency for Health Care Policy and Research.

Mosteller, F., Gilbert, J.P., & McPeek, B. (1980). Reporting standards and research strategies for controlled trials: Agenda for the editor. Controlled Clinical Trials, 1, 37-58.

Mosteller, F., Light, R. J., & Sachs, J. A. (1996). Sustained inquiry in education: Lessons from skill grouping and class size. Harvard Educational Review, 66, 797-842.

Mulford, H. A., Ledolrer, J., & Fitzgerald, J. L. (1992). Alcohol availability and consumption: Iowa sales data revisited. Journal of Studies on Alcohol, 53, 487-494.

Mulkay, M. (1979). Science and the sociology of knowledge. London: Allen & Unwin.

Mullin, B. (1993). Advanced BASIC meta-analysis. Hillsdale, NJ: Erlbaum.

Mumford, E., Schlesinger, H. J., Glass, G. V., Patrick, C., & Cuerdon, T. (1984). A new look at evidence about reduced cost of medical utilization following mental health treatment. American Journal of Psychiatry, 141, 1145-1158.

Murdoch, J. C., Singh, H., & Thayer, M. (1993). The impact of natural hazards on housing values: The Loma Prieta earthquake. Journal of the American Real Estate and Urban Economics Association, 21, 167-184.

Murnane, R. J., Newstead, S., & Olsen, R. J. (1985). Comparing public and private schools: The puzzling role of selectivity bias. Journal of Business and Economic Statistics, 3, 23-35.

Murray, C. (1984). Losing ground: American social policy, 1950-1980. New York: BasicBooks.

Murray, D. M. (1998). Design and analysis of group-randomized trials. New York: Oxford University Press.

Murray, D. M., & Hannan, P. J. (1990). Planning for the appropriate analysis in school-based drug-use prevention studies. Journal of Consulting and Clinical Psychology, 58, 458-468.

Murray, D. M., Hannan, P. J., & Baker, W. L. (1996). A Monte Carlo study of alternative responses to intraclass correlation in community trials: Is it ever possible to avoid Cornfield's penalties? Evaluation Review, 20, 313-337.

Murray, D. M., McKinlay, S. M., Martin, D., Donner, A. P., Dwyer, J. H., Raudenbush, S. W., & Graubard, B. I. (1994). Design and analysis issues in community trials. Evaluaion Review, 18, 493-514.

Murray, D. M., Moskowitz, J. M., & Dent, C. W. (1996). Design and analysis issues in community-based drug abuse prevention. American Behavioral Scientist, 39, 853-867.

Muthen, B. O. (1994). Multilevel covariance structure analysis. Sociological Methods and Research, 22, 376-398.

Muthen, B. O., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing-completely at random. Psychometrika, 52, 431-462.

Narayanan, V. K., & Nath, R. (1982). A field test of some attitudinal and behavioral consequences of flexitime. Journal of Applied Psychology, 67, 214-218.

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1979). The Belmont report: Ethical principles and guidelines for the protection of human subjects of research (OPRR Report; FR Doc. No. 79-12065). Washington, DC: U.S. Government Printing Office.

National Institutes of Health. (1994). NIH guidelines on the inclusion of women and minorities as subjects in clinical research, 59 Del. Reg. 14, 508 (Document No. 94-5435).

Naylor, R. H. (1989). Galileo's experimental discourse. In D. Gooding, T. Pinch, & S. Schaffer (Eds.), The uses of experiment: Studies in the natural sciences (pp. 117-134). Cambridge, England: Cambridge University Press.

Neal-Schuman Publishers. (Eds.). (1980). National Directory ofMental Health. New York: Wiley.

Nesselroade, J. R., Stigler, S. M., & Baltes, P. B. (1980). Regression to the mean and the study of change. Psychological Bulletin, 88, 622-637.

Neter, J., Wasserman, W., & Kutner, M. H. (1983). Applied linear regression models. New York: Gardner Press.

Neustrom, M. W., & Norton, W. M. (1993). The impact of drunk driving legislation in Louisiana. Journal of Safety Research, 24, 107-121.

Newbold, P., Agiakloglou, C., & Miller, J. (1994). Adventures with ARIMA software. International Journal of Forecasting, 10, 573-581.

Newhouse, J. P. (1993). Free for all? Lessons from the RAND Health Insurance Experiment. Cambridge, MA: Harvard University Press.

Newhouse, J.P., & McClellan, M. (1998). Econometrics in outcomes research: The use of instrumental variables. Annual Review of Public Health, 19, 17-34.

Nicholson, R. A., & Berman, J. S. (1983). Is follow-up necessary in evaluating psychotherapy? Psychological Bulletin, 93, 261-278.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. Psychological Methods, 5, 241-301.

Nietzel, M. T., Russell, R. L., Hemmings, K. A., & Gretter, M. L. (1987). Clinical significance of psychotherapy for unipolar depression: A meta-analytic approach to social comparison. Journal of Consulting and Clinical Psychology, 55, 156-161.

Notz, W. W., Staw, B. M., & Cook, T. D. (1971). Attitude toward troop withdrawal from Indochina as a function of draft number: Dissonance or self-interest? Journal of Personality and Social Psychology, 20, 118-126.

Nunnally, J. C., & Bernstein, I. H. (1994). Psychometric theory (3rd ed.). New York: McGraw-Hill.

Nurco, D. N., Robins, L. N., & O'Donnel, J. A. (1977). Locating respondents. In L. D. Johnston, D. N. Nurco, & L. N. Robins (Eds.), Conducting follow-up research on drug treatment programs (NIDA Treatment Program Monograph Series, No.2, pp. 71-84). Rockville, MD: National Institute on Drug Abuse.

Nuremberg Code. (1949). Trials of war criminals before the Nuremberg Military Tribunals under Control Council Law No. 10 (Vol. 2). Washington, DC: U.S. Government Printing Office.

Oakes, D., Moss, A. J., Fleiss, J. L., Bigger, J. T., Jr., Therneau, T., Eberly, S. W., McDermott, M. P., Manatunga, A., Carleen, E., Benhorn, J., and The Multicenter Diltiazem Post-Infarction Trial Research Group. (1993). Use of compliance measures in an analysis of the effect of Diltiazem on mortality and reinfarction after myocardial infarction. Journal of the American Statistical Association, 88, 44-49.

O'Carroll, P. W., Loftin, C., Waller, J. B., McDowall, D., Bukoff, A., Scott, R. O., Mercy, J. A., & Wiersema, B. (1991). Preventing homicide: An evaluation of the efficacy of a Detroit gun ordinance. American Journal of Public Health, 81, 576-581.

O'Connor, F. R., Devine, E. C., Cook, T. D. & Curtin, T. R. (1990). Enhancing surgical nurses' patient education. Patient Education and Counseling, 16, 7-20.

Okene, J. (Ed.). (1990). Adoption and maintenance of behaviors for optimal health. New York: Academic Press.

Oldroyd, D. (1986). The arch of knowledge: An introductory study of the history of the philosophy and methodology of science. New York: Methuen.

Olds, D. L., Eckenrode, D., Henderson, C. R., Kitzman, H., Powers, J., Cole, R., Sidora, K., Morris, P., Pettitt, L. M., & Luckey, D. (1997). Long-term effects of home visitation on maternal life course and child abuse and neglect. Journal of the American Medical Association, 278, 637-643.

Olds, D. L, Henderson, C. R., Cole, R., Eckenrode, J., Kitzman, H., Luckey, D., Pettitt, L. M., Sidora, K., Morris, P., & Powers, J. (1998). Long-term effects of nurse home visitation on children's criminal and antisocial behavior: 15-year follow-up of a randomized controlled trial. Journal of the American Medical Association, 280, 1238-1244.

Olds, D., Henderson, C. R., Kitzman, H., & Cole, R. (1995). Effects of prenatal and infancy nurse home visitation on surveillance of child maltreatment. Pediatrics, 95, 365-372.

O'Leary, K. D., Becker, W. C., Evans, M. B., & Saudargas, R. A. (1969). A token reinforcement program in a pnblic school: A replication and systematic analysis. Journal of Applied Behavior Analysis, 3, 3-13.

O'Leary, K. D., & Borkovec, T. D. (1978). Conceptual, methodological, and ethical problems of placebo groups in psychotherapy research. American Psychologist, 33, 821-830.

Ones, D. S., Viswesvaran, C., & Schmidt, F. L. (1993). Comprehensive meta-analysis of integrity test validities: Findings and implications for personnel selection and theories. Of job performance [Monograph]. Journal of Applied Psychology, 78, 679-703.

Orne, M.T. (1959). The nature of hypnosis: Artifact and essence. Journal of Abnormal and Social Psychology, 58, 277-299.

Orne, M. T. (1962). On the social psychology of the psychological experiment. American Psychologist, 17, 776-783.

Orne, M. T. (1969). Demand characteriStics and the concept of quasi-controls. In R. Rosenthal & R. L. Rosnow (Eds.), Artifact in behavioral research (pp. 143-179). New York: Academic Press.

Orr, L. L. (1999). Social experiments: Evaluating public programs with experimental methods. Thousand Oaks, CA: Sage.

Orr, L. L., Johnston, T., Montgomery, M., & Hojnacki, M. (1989). Design of the Washington Self-Employment and Enterprise Development (SEED) Demonstration. Bethesda, MD: Abt/Battell Memorial Institute.

Orwin, R. G. (1983). A fail-safe N for effect size in meta-analysis. Journal of Educational Statistics, 8, 157-159.

Orwin, R. G. (1984). Evaluating the life cycle of a product warning: Saccharin and diet soft drinks. Evaluation Review, 8, 801-822.

Orwin, R. G. (1994). Evaluating coding decisions. In H. Cooper & L. V. Hedges (Eds.), The handbook of research synthesis (pp. 140-162). New York: Russell Sage Foundation.

Orwin, R. G., & Cordray, D. S. (1985}. Effects of deficient reporting on meta-analysis: A conceptual framework and reanalysis. Psychological Bulletin, 8S, 185-193.

Orwin, R. G., Cordray, D. S., & Huebner, R. B. (1994). Judicious application of randomized designs. In K. J. Conrad (Ed.), Critically evaluating the role of experiments (pp. 73-86). San Frandsen: Jossey-Bass.

Ostrom, C. W. (1990). Time series analysis: Regression techniques (2nd ed.). Thousand Oaks, CA: Sage.

Ottenbacher, K. J. (1986). Reliability and accuracy of visually analyzing graphed dara from single-subject designs. American Journal of Ocwpational Therapy, 40, 464-469.

Overall, J. E., & Woodward, J. A. (1977). Nonrandom assignment and the analysis of covariance. Psychlogical Bulletin, 84, 58&--594.

Overton, R. C. (1998). A comparison of fixed-effects and mixed (random-effects) models for meta-analysis tests of moderator variable effects. Psychological Methods. 3, 354-379.

Ozminkowski, R. J., Wortman, P.M., & Roloff, D. W. (1989). Inborn/outborn status and neonatal survival: A meta-anaJysis of non-randomized studies. Statistics in Medicine, 7, 1207-1221.

Page, E. B. (1958). Teacher comments and student performance: A seventy-four classroom experiment in school motivation. Jaumal of Educational Psychology, 49, 173-181.

Palmer, C. R., & Rosenberger, W. F. (1999). Ethics and practice: Alternarive designs for Phase Ill randomized clinical trials. Controlled Clinical Trials, 20, 172-186.

Patterson, G. R. (1986). Performance models for antisocial boys. American Psychologist, 41, 432-444.

Patron, M, Q. (1980). Qualitatve evaluation methods. Beverly Hills, CA: Sage Publications.

Paulos, J. A. (1988). Innumeracy: Mathematical illiteracy and its consequences. New York: Hill & Wang.

Pearl, J. 12000). Causality: Models, reasoning and inference. Cambridge, England: Cambridge University Press.

Pearhnan, S., Zweben, A., & Li, $, (1989). The comparability of solicited versus clinic subjects in alcohol treatment research. British Journal of Addictions, 84, 523-532.

Pearson, K. {1904). Report on certain entiric fever inoculation statistics. British Medical

Journal, 2, 1243-1246.

Peirce, C. S., & Jastrow, J. (1884).On small diiferences of sensation. Memoirs of the National Academy of Sciences, 3, 75-83.

Pelz, D. C., & Andrews, F. M. (1964). Detecting causal priorities in panel study data. American Sociological Review, 29, 836-848.

Permutt, T. {1990). Testing for imbalance of covariates in controlled experiments. Statistics in Medicine, 12, 1455-1462.

Perng, S, S, (1985). Accounts receivable treatments study. In R. F. Boruch & W, Wothke (Eds.), Randmization and field experimentation {pp. 55--62). San Francisco: Jossey-Bass.

Petersen, W. M. (1999, April 20). Economic status quo [Letter ro the editor]. New York Times, p. A18.

Petty. R. E., Fabrigar, L R., Wegener, D. T., & Priester, J. R. (1996). Understanding data when interactions are present or hypothesized. Psychological Science, 7, 247-252.

Pfungst, O. (1911). Clever Hans (the horse ofMr. Von Osten). New York: Henry Holt.

Phillips, D. C (1987). Validity in qualitative research: Why the worry about warrant will not wane. Education and Urban Society, 20, 9-24.

Phillips, D.C (1990). Postpositivistic science: Myths and realities. In E. G. Guba (Ed.), The paradigm dialog (pp. 31-45). Newbury Park, CA: Sage.

Pigott, T. D. (1994). Methods for handling missing data in research syntheses. In H. Cooper & L. V. Hedges (Eds.), The handbook of research synthesis (pp. 163-175). New York: Russell Sage Foundation.

Pinch, T. (1986). Confronting nature. Dordrecht, Holland: Reidel.

Pirie, P. L, Thomson, M, A., Mann, S. L., Peterson, A. V., Murray, M., Flay, B. R., & Best, J. A. {1989). Tracking and attrition in longitudinal school-based smoking prevention research. Preventive Medicine, 18, 249-156.

Platt, J. R. (1964). Strong inference. Science, 146, 347-353.

Plomin, R., & Daniels, D. (1987). Why are children from the same family so different from one another? Behavioral and Brain Sciences, 10, 1-60.

Pocock, S.J. (1983). Clinirol trials: A practical approach. Chichester, England: Wiley.

Pocock, S. J., & Simon, R. (1975). Sequential treatment assignment with balancing for prognostic factors in the controlled clinical triaL Bicmretrics, 31, 103-115.

Polanyi, M. (1958}. Personal knowledge: Toward a post-critical philosophy. London: Routledge & kegan Paul.

Polsby, N. W. (1984). Political innovation in America: The politics of policy tnitiation. New Haven, CT: Yale Universiry Press.

Popper, K. R. {1959}. The logic of scientific discovery. New York: Basic Books.

Population Council. (1986). An experimental study of the efficiency and effectiveness of an IUD insertion and back-up component (English summary; Report No. PC PES86). Lima, Peru; Author.

Potvin, L., & Campbell, D. T. (1996). Exposure opportunity, the experimental paradigm and the case-control study. Unpublished manuscript.

Pound, C. R., Partin, A. W., Eisenberger, M.A., Chan, D. W., Pearson,J. D., & Walsh, P. C. (1999). Natural history of progression after PSA elevation following radical prostatectomy. Journal of the American Medical Asociation, 281, 1591-1597.

Powers, K. L., & Anglin, M. D. (1993). Cumulative versus stabilizing effects of methadone maintenance: A quasi-experimental study using longitudinal self-report data. Evaluation Review, 17, 243-270.

Powers, E., & Witmer, H. (1951}. An experiment in the prevention of delinquency. New York: Columbia University Press.

Premack, S. L, & Hunter, J. E. (1988). Individual unionization decisions. Psychological Bulletin, 103, 223-234.

Prentice, D. A., & Miller, D. T. (1992}. When small effects are impressive. Psychological Bulletin, 112, 16Q-164.

Presby, S. (1978). Overly broad categories obscure important differeinces between therapies. American Psychologist, 33, 514-515.

Pressman, J. L, & Wildavsky, A. (1984). Implementation (3rd ed.). Berkeley: University of California Press.

Project MATCH Research Group (1993). Project MATCH: Rationale and methods for a multisite clinical trial matching patients to alcoholism treatment. Alcoholism: Clinical and Experimental Research, 17, 1130-1145.

Protection of Human Subjects, Title 45 C.P.R. Part 46, Subparts A-D (Revised 1991).

Psaty, B. M., Koepsell, T. D., Lin, D., Weiss, N. S., Siscovick, D. S., Rosendaal, F. R., Pahor, M., & Furberg, C. D. (1999). Assessment and control for confounding by indication in observational studies. Journal of the American Geriatric Society, 47, 749-754.

Puma, M. J., Burstein, N. R., Merrell, K., & Silverstein, G. (1990). Evaluation of the Food Stamp Employment and Training Program. Final report: Volume I. Bethesda, MD: Abt.

Quine, W. V. (1951). Two dogmas of empiricism. Philosophical Review, 60, 20-43.

Quine, W. V. (1969). Ont ological relativity and other essays. New York: Columbia University Press.

Quirk, P. J. (1986). Public policy [Review of Political Innovation in America and Agendas, Alternatives, and Public Policies]. Journal of Policy Analysis and Management, 6, 607-613.

Ralston, D. A., Anthony, W. P., & Gustafson, D. J. (1985). Employees may love flextime, but what does it do to the organization's productivity? Journal of Applied Psychology, 70, 272-279.

Ranstam, J., Buyse, M., George, S. L., Evans, S., Geller, N. L., Scherrer, B., Lasaffre, E., Murray, G., Edler, L., Hutton, J. L., Colton, T., & Lachenbruch, P. (2000). Fraud in medical research: An international survey of biostatisticians. Controlled Clinical Trials, 21, 415-427.

Raudenbush, S. W. (1994). Random effects models. In H. Cooper & L. V. Hedges (Eds.), The handbook of research synthesis (pp. 301-321). New York: Russell Sage Foundation.

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized design. Psychological Methods, 2, 173-185.

Raudenbush, S. W., Becker, B. J., & Kalaian, H. (1988). Modeling multivariate effect sizes. Psychological Bulletin, 103, 111-120.

Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. Psychological Methods, 5, 199-213.

Raudenbush, S. W., & Willms, J.D. (Eds.). (1991). Schools, classrooms, and pupils: International studies of schooling from a multilevel perspective. San Diego, CA: Academic Press.

Raudenbush, S. W., & Willms, J.D. (1995). The estimation of school effects. Journal of Educational and Behavioral Statistics, 20, 307-335.

Rauma, D., & Berk, R. A. (1987). Remuneration and recidivism: The long-term impact of unemployment compensation on ex-offenders. Journal of Quantitative Criminology, 3, 3-27.

Ray, J. W., & Shadish, W. R. (1996). How interchangeable are different estimators of effect size? Journal of Consulting and Clinical Psychology, 64, 1316-1325. (See also "Correction to Ray and Shadish (1996)," Journal of Consulting and Clinical Psychology, 66, 532, 1998)

Reding, G. R., & Raphelson, M. (1995). Around-the-clock mobile psychiatric crisis intervention: Another effective alternative to psychiatric hospitalization. Community Mental Health Journal, 31, 179-187.

Reed, J. G., & Baxter, P. M. (1994). Using reference databases. In H. Cooper & L. V. Hedges (Eds.), The handbook of research synthesis (pp. 57-70). New York: Russell Sage Foundation.

Rees, A. (1974). The graduated work incentive experiment: An overview of the labor-supply results. Journal of Human Resources, 9, 158-180.

Reichardt, C. S. (1985). Reinterpreting Seaver's (1973) study of teacher expectancies as a regression artifact. Journal of Educational Psychology, 77. 231-236.

Reichardt, C. S. (1991}. Comments on "The application of time serieS methods to moderate span longitudinal data." In L. M. Collins & J. L. Horn (Eds.), Best methods for the analysis of change: Recent

advances, unanswered questions, future directions (pp. 88-91). Washington, DC: American Psychological Asociation.

Reichardt, C. S. (2000). A typology of strategies for ruling out threats to validity. In L. Bickman (Ed.), Research design: Donald Campbell's legacy (Vol. 2, pp. 89-115). Thousand Oaks, CA: Sage.

Reichardt, C. S., & Gollob, H. E (1986). Satisfying the constraints of causal modeling. In W. M. K. Trochim (Ed.), Advances in quasi-experimental design and analysis (pp. 91-107). San Francisco: Jossey·Bass.

Reichardt, C. S., & Gollob, H. F. (1987, October). Setting limits on the bias due to omitted uariables. Paper presented at the meeting of the Society of Multivariate Experimental Psychology, Denver, CO.

Reichardt, C. S., & Gollob, H. F. (1997). When confidence intervals should be used instead of statistical tests, and vice versa. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (FAs.), What if there were no significance tests? (pp. 259-284). Hillsdale, KJ: Erlhaum.

Reichardt, C. S., & Rallis, S. F. (Eds.). (1994). The qualitative-quantitative debate: New perspectives. San Francisco: Jossey-Bass.

Reichardt, C. S., Trochim, W. M. K., & Cappelleri, J. C. (1995). Reports of the death of the regression-discontinuity design are greatly exaggerated. Evaluation Review, J9, 39-63.

Reicken, H. W., Boruch, R. F.; Campbell, D. T., Caplan, N., Glennan, T. K., Pratt, J. W., Rees, A., & Williams, W. (1974). Social experimentation: A method for planning and evaluating social intervention. New York: Academic Press.

Reinsel, G. (1993). Elements of multivariate time series analysis. New York: Springer-Verlag.

Rescher, N. (1969). Introduction to value theory. Engle wood Cliffs, NJ: Prentice-Hall.

Revelle, W., Humphreys, M. S., Simon, L., & Gililland, K. (1980). The interactive effect of personality, time of day, and caffeine: A rest of the arousal model. Journal of Experimental Psychology: General, 109, 1-31.

Reynolds, A. J., & Temple, J. A. (1995). Quasi-experimental estimates of the effects of a preschool intervention: Psychometric and econometric comparisons. Evaluation Review, 19,347-373.

Reynolds, K. D., & West, S. G. (1987). A multiplist strategy for strengthening nonequivalent control group designs. Evaluation Review, 11, 691-714.

Rezmovic, E. L., Cook, T. J., & Dobson, L. D. (1981). Beyond random assignment: Factors affecting evaluation integrity. Evaluation Review, 5, 51-67.

Ribisl, K. M., Walton, M.A., Mowbray, C. T., Luke, D. A., Davidson, W. S., & Bootsmiller, B. J. (1996). Minimizing participant attrition in panel studies through the use of effective retention and tracking strategies: Review and recommendations. Evaluation and Program Planning, 19, 1-25.

Rindskopf, D. (1986). New developments in selection modeling for quasi-experimentation. In W. M. K. Trochiin (Ed.), Advances in quasi-experimental design and analysis (pp. 79-89). San Francisco: Jossey-Bass.

Rindskopf, D. (2000). Plausible rival hypotheses in measurement, design, and scientific theory. In L. Bickman (Ed.), Research design: Donald Campbell's legacy (Vol. 1, pp. 1-12). Thousand Oaks, CA: Sage.

Rindskopf, D.M. (1981). Structural equation models in analysis of nonexperimental data. In R. F. Boruch, P. M. Wortman, & D. S. Cordray (Eds.), Reanalyzing program evaluations (pp. 163-193). San Francisco: Jossey-Bass.

Rivlin, A.M., & Timpane, P.M. (1975). (Eds.). Planned variation in education. Washington, DC: Brookings.

Robbins, H., & Zhang, C.-H. {i988). Estimating a treatment effect under biased sampling. Proceedings of the National Academy of Science, USA, 85, 3670-3672.

Robbins, H., & Zhang, C.-H. (1989). Estimating the superiority of a drug to a placebo when all and only those patientS at risk are treated with the drug. Proceedings of the National Academy of Science, USA, 86, 3003-3005.

Robbins, H., & Zhang, C.-H. (1990). Estimating a treatment effect under biased allocation. (Working paper.) New Brunswick, NJ: Rutgers University, Institute of Biostatistics and Department of Statistics.

Roberts, J. V., & Gebotys, R. J. (1992). Reforming rape laws: Effects of legislative change in Canada. Law and Human Behavior, 16, 555-573.

Robertson, T. S., & Rossite, J. R. (1976). Short-run advertising effects on children: A field study. Journal of Marketing Research, 8, 68-70.

Robins, J. M. (1998). Correction for non-compliance in equivalence trials. Statistics in Medicine, 17, 269-302...

Robins, J. M. (2000). Marginal structural models versus structural nested models as tools for causal inference. In M. E. Halloran & D. Berry (Eds.), Statistical models in epidemiology, the environment, and clinical trials (pp. 95-133). New York: Springer-Verlag.

Robins, J. M., & Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. Epidemiology, 3, 143-155.

Robins, J. M., & Greenland, S. (1996). Comment. Journal of the American Statistical Association, 91, 456-458.

Robins, J. M., Greenland, S., & Hu, F.-C. (1999). Estimation of the causal effect of a time- varying exposure on the marginal mean of a repeated binary outcome (with discussion). Journal of the American Statistical Association, 94, 687-712.

Robinson, A., Bradley, R. D., & Stanley, T. D. (1990). Opportunity to achieve: Identifying and serving mathematically talented black students. Contemporary Educational Psychology, 15, 1-12.

Robinson, A., & Stanley, T. D. (1989). Teaching to talent: Evaluating an enriched and accelerated mathematics program. Journal of the Education of the Gifted, 12, 253-267.

Robinson, L.A., Berman, J. S., & Neimeyer, R. A. (1990). Psychotherapy for the treatment of depression: A comprehensive review of controlled outcome research. Psychological Bulletin, 108, 30-49.

Rockette, H. E. (1993). What evidence is needed to link lung cancer to second-hand smoke? Chance: New Directions for Statistics and Computing, 6, 15-18.

Roese, N. J., & Jamieson, D. W. (1993). Twenty years of bogus pipeline research: A critical review and meta-analysis. Psychological Bulletin, 114, 363-375.

Roethlisberger, F. S., & Dickson, W. J. (1939). Management and the worker. Cambridge, MA: Harvard University Press.

Rogers, J. L., Howard, K. I., & Vessey, J. T. (1993). Using significance tests to evaluate equivalence between two experimental groups. Psychological Bulletin, 113,553-565.

Rogers, P. J., Hacsi, T. A., Petrosino, A., & Huebner, T. A. (Ed.s.). (2000). Program theory in evaluation: Challenges and opportunities. San Francisco: Jossey-Bass.

Rogosa, D. (1980). A critique of cross-lagged correlation. Psychological Bulletin, 88,245-258.

Rogosa, D. (1988). Myths about longitudinal research. In K. W. Schaie, R. T. Campbell, W. Meredith, & S. C. Rawlings (Eds.), Methodological issues in aging research (pp. 171-210). New York: Springer.

Rosch, E. H. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (Eds.), Cognition and categorization (pp. 27-48). Hillsdale, NJ: Erlbaum.

Rosenbaum, P.R. (1984). From association to causation in observational studies: The role of tests of strongly ignorable treatment assumptions. Journal of the American Statistical Association, 79, 41-48.

Rosenbaum, P.R. (1986). Dropping out of high school in the United States: An observational study. Journal of Educational Statistics, 11, 207-224.

Rosenbaum, P.R. (1987). Sensitivity analysis for certain permutation tests in matched observational studies. Biometrika, 74, 13-26.

Rosenbaum, P. R. (1988). Sensitivity analysis for matching with multiple controls. Biometrika, 75, 577-581.

Rosenbaum, P. R. (1989). Sensitivity analysis for matched observational studies with many brdered treatments. Scandinavian Journal of Statistics, 16, 227-236.

Rosenbaum, P. R. (1991a). Discussing hidden bias in observational studies. Annals of Internal Medicine, 115, 901-905.

Rosenbaum, P.R. (1991b). Sensitivity analysis for matched case-control studies. Biometrics, 47, 87-100.

Rosenbaum, P.R. (1993). Hodges-Lehmann point estimates of treatment effect in observational studies. Journal of the American Statistical Association, 88, 1250-1253.

Rosenbaum, P.R. (1995a). Observational studies. New York: Springer-Verlag.

Rosenbaum, P. R. (1995b). Quantiles in nonrandom samples and observational studies. Journal of the American Statistical Association, 90, 1424-1431.

Rosenbaum, P. R. (1996a). Comment. Journal of the American Statistical Association, 91, 465-468.

Rosenbaum, P.R. (1996b). Observational studies and nonrandomized experiments. In S. Ghosh & C. R. Rao (Eds.), Handbook of statistics (Vol. 13, pp. 181-197). Amsterdam: Elsevier Science.

Rosenbaum, P. R. (1998). Multivariate matching methods. In S. Kotz, N. L. Johnson, L. Norman, & C. B. Read (Eds.), Encyclopedia of statistical sciences (Update Volume 2, pp. 435-438). New York: Wiley.

Rosenbaum, P.R. (1999a). Choice as an alternative to control in observational studies. Statistical Science, 14, 259-304.

Rosenbaum, P.R. (1999b). Using quantile averages in matched observational studies. Applied Statistics, 48, 63-78.

Rosenbaum, P. R. (in press). Observational studies. In N. J. Smelser & P. B. Baltes (Eds.), International encyclopedia of social and behavioral sciences. Oxford, England: ElsevierScience.

Rosenbaum, P.R., & Krieger, A. (1990). Sensitivity analysis for matched case-control studies. Journal of the American Statistical Association, 85, 493-498.

Rosenbaum, P. R., & Rubin, D. (1984). Reducing bias in observational studies using subclassification on the propensity score. Journal of the American Statistical Association, 79,516-524.

Rosenberg, M., Adams, D. C., & Gurevitch, J. (1997). Meta Win: Statistical software for meta-analysis with resampling tests. (Available from Sinauer Associates, Inc., P.O. Box 407, Sunderland, MA 01375-0407)

Rosenberg, M. J. (1969). The conditions and consequences of evaluation apprehension. In R. Rosenthal & R. L. Rosnow (Eds.), Artifact in behavioral research (pp. 279-349). New York: Academic Press.

Rosenberger, W. F. (1999). Randomized play-the-winner clinical trials: Review and recommendations. Controlled Clinical Trials, 20, 328-342.

Rosenthal, M. C. (1994). The fugitive literature. In H. Cooper & L. V. Hedges (Eds.), The handbook of research synthesis (pp. 85-94). New York: Russell Sage Foundation.

Rosenthal, R. (1956). An attempt at the experimental induction of the defense mechanism of projection. Unpublished doctoral dissertation, University of California, Los Angeles.

Rosenthal, R. (1966). Experimenter effects in behavioral research. New York: Appleton-Century-Crofts.

Rosenthal, R. (1973a). The mediation of Pygmalion effects: A four-factor theory. Papua New Guinea Journal of Education, 9, 1-12.

Rosenthal, R. (1973b). On the social psychology of the self-fulfilling prophecy: Further evidence for Pygmalion effects and their mediating mechanisms. New York: MSS Modular Publication, Module 53.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. Psychological Bulletin, 86, 638-641.

Rosenthal, R. (1986). Meta-analytic procedures and the nature of replication: The ganzfeld debate. Journal of Parapsychology, 50, 315-336.

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.), The handbook of research synthesis (pp. 231-244). New York: Russell Sage Foundation.

Rosenthal, R., & Rosnow, R. L. (1991). Essentials of behavioral research: Methods and data analysis. New York: McGraw-Hill.

Rosenthal, R., & Rubin, D. B. (1978). Interpersonal expectancy effects: The first 345 studies. Behavioral and Brain Sciences, 3, 377-386.

Rosenthal, R., & Rubin, D. B. (1986). Meta-analytic procedures for combining studies with multiple effect sizes. Psychological Bulletin, 99, 400-406.

Rosenthal, R., & Rubin, D. B. (1994). The counternull value of an effect size: A new statistic. Psychological Science, 5, 329-334.

Rosenzweig, S. (1933). The experimental situation as a psychological problem. Psychological Review, 40, 337-354.

Rosnow, R. L., & Rosenthal, R. (1989). Definition and interpretation of interaction effects. Psychological Bulletin, 105, 143-146.

Rosnow, R. L., & Rosenthal, R. (1996). Contrasts and interaction effects redux: Five easy pieces. Psychologiwl Science, 7, 253-257.

Rosnow, R. L., & Rosenthal, R. (1997). People studying people: Artifacts and ethics in behavioral research. New York: Freeman.

Ross, A. S., & Lacey, B. (1983). A regression discontinuity analysis of a remedial education programme. Canadian Journal of Higher Education, 13, 1-15.

Ross, H. L. (1973). Law, science and accidents: The British Road Safety Act of 1967.Journal of Legal Studies, 2, 1-75.

Ross, H. L., Campbell, D. T., & Glass, G. V. (1970). Determining the social effects of a legal reform: The British "breathalyser" crackdown of 1967. American Behavioral Scientist, 13, 493-509.

Rossi, P. H. (1987). The iron law of evaluation and other metallic rules. In J. Miller & M. Lewis (Eds.), Research in social problems and public policy (Vol. 4, pp. 3-20). Green- wich, CT: JAI Press.

Rossi, P. H. (1995). Doing good and getting it right. In W. R. Shadish, D. L.Newman, M.A. Scheirer, & C. Wye (Eds.), Guiding principles for evaluators (pp. 55-59). San Francisco: Jossey-Bass.

Rossi, P. H., Berk, R. A., & Lenihan, K. J. (1980). Money, work and crime: Some experimental findings. New York: Academic Press.

Rossi, P. H., & Freeman, H. E. (1989). Evaluation: A systematic approach (4th ed). Thousand Oaks, CA: Sage.

Rossi, P. H., Freeman, H. E., & Lipsey, M. W. (1999). Evaluation: A systematic approach (6th ed.). Thousand Oaks, CA: Sage;

Rossi, P. H., & Lyall, K. C. (1976). Reforming public welfare. New York: Russell Sage.

Rossi, P. H., & Lyall, K. C. (1978). An overview evaluation of the NIT Experiment. In T. D. Cook, M. L. DelRosario, K. M. Hennigan, M. M. Mark, & W. M. K. Trochim (Eds.), Evaluation studies review annual (Volume 3, pp. 412-428). Newbury Park, CA: Sage.

Rossi, P. H., & Wright, J.D. (1984). Evaluation research: An assessment. Annual Review of Sociology, 10, 331-352.

Rossi, P. H., Wright, J.D., & Anderson, A. B. (Eds.). (1983). Handbook of survey research. San Diego, CA: Academic Press.

Rothman, K. (1986). Modern epidemiology. Boston: Little Brown.

Rouanet, H. (1996). Bayesian methods for assessing importance of effects. Psychological Bulletin, 119, 149-158.

Rouse, C. E. (1998). Private school vouchers and student achievement: An evaluation of the Milwaukee Parental Choice Program. Quarterly Journal of Economics, 113, 553-602.

Rowe, P.M. (1999). What is all the hullabaloo about endostatin? Lancet, 353, 732.

Rozeboom, W. W. (1960). The fallacy of the null hypothesis significance test. Psychological Bulletin, 57, 416-428.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology, 66, 688-701.

Rubin, D. B. (1977). Assignment to treatment group on the basis of a covariate. Journal of Educational Statistics, 2, 1-26.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. Annals of Statistics, 6, 34-58.

Rubin, D. B. (1986). Comment: Which ifs have causal answers. Journal of the American Statistical Association, 81, 961-962.

Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys. New York: Wiley.

Rubin, D. B. (1990). A new perspective. In K. W. Wachter & M. L. Straf (Eds.), The future of meta-analysis (pp. 155-165). New York: Russell Sage Foundation.

Rubin, D. B. (1991). Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism. Biometrics, 47, 1213-1234.

Rubin, D. B. (1992a). Clinical trials in psychiatry: Should protocol deviation censor patient data? A comment. Neuropsychopharmacology, 6, 59-60.

Rubin, D. B. {1992b). Meta-analysis: Literature synthesis or effect-size surface estimation? Journal of Educational Statistics, 17, 363-374.

Rubin, D. B. {1997). Estimating causal effects from large data sets using propensity scores. Annals of Internal Medicine, 127, 757-763.

Rubin, D. B., & Thomas, N. {1996). Matching using estimated propensity scores: Relating theory to practice. Biometrics, 52, 249-264.

Rubin, D. B., & Thomas, N. (2000). Combining propensity score matching with additional adjustments for prognostic covariates. Journal of the American Statistical Association, 95, 573-585.

Rubins, H. B. (1994). From clinical trials to clinical practice: Generalizing from participant to patient. Controlled Clinical Trials, 15, 7-10.

Rudd, P., Ahmed, S., Zachary, V., Barton, C., & Bonduelle, D. (1990). Improved compliance measures: Applications in an ambulatory hypertensive drug trial. Clinical Pharmacological Therapy, 48, 676-685.

Ryle, G. (1971). Collected papers (Vol. 2). London: Hutchinson.

Sackett, D. L. (1979). Bias in analytic research. Journal of Chronic Diseases, 32, 51-63.

Sackett, D. L. (2000). Why randomized controlled trials fail but needn't: 1. Failure to gain "coal-face" commitment and to use the uncertainty principle. Canadian Medical Association Journal, 162, 1311-1314.

Sackett, D. L., & Haynes, R. B. (1976). Compliance with therapeutic regimens. Baltimore: Johns Hopkins University Press.

Sackett, D. L., & Hoey, J. (2000). Why randomized controlled trials fail but needn't: A new series is launched. Canadian Medical Association Journal, 162, 1301-1302.

Sacks, H. S., Chalmers, T. C., & Smith, H. (1982). Randomized versus historical controls for clinical trials. American Journal of Medicine, 72, 233-240.

Sacks, H. S., Chalmers, T. C., & Smith, H. (1983). Sensitivity and specificity of clinical trials: Randomized v historical controls. Archives of Internal Medicine, 143,753-755.

St. Pierre, R. G., Cook, T. D., & Straw, R. B. (1981). An evaluation of the Nutrition Education and Training Program: Findings from Nebraska. Evaluation and Program Planning, 4, 335-344.

St. Pierre, R. G., Ricciuti, A., & Creps, C. (1998). Summary of state and local Even Start evaluations. Cambridge, MA: Abt.

Sales, B. D., & Folkman, S. (Eds.). (2000). Ethics in research with human participants. Washington, DC: American Psychological Association.

Salmon, W. C. (1984). Scientific explanation and the causal structure of the world. Princeton, NJ: Princeton University Press.

Salmon, W. C. (1989). Four decades of sciefitific explanation. Minneapolis: University of Minnesota Press.

Salner, M. (1989). Validity in human science research. In S: Kvale (Ed.), Issues ofvalidity in qualitative research (pp. 47-71). Lund, Sweden: Studentlitteratur.

Sampson, R. J., Raudenbush, S. W., & Earls, F. (1997). Neighborhoods and violent crime: A multilevel study of collective efficacy. Science, 277, 918-924.

Sanchez-Meca, J., & Marin-Martinez, F. (1998). Weighting by inverse variance or by sample size in meta-anaiysis: A simulation study. Educational and Psychological Measurement, 58, 211-220.

Sarason, S. B. (1978). The nature of problem solving in social action. American Psychologist, 33, 370-380.

Saretsky, G. (1972). The OEO PC experiment and the John HenRy effect. Phi Delta Kappan, 53, 579-581.

Sargent, D. J., Sloah, J. A., & Cha, S. S. (1999). Sample size and design considerations for Phase ll clinical trials with correlated observations. Controlled Clinical Trials. 19, 242-252. '

Saunders, L. D., lrwig, L. M., Gear, J. S., & Ramushu, D. L. (1991). A randomized controlled trial of compliahce improving strategies in Soweto hypertensives. Medical Care, 29, 669-678.

Sayrs, L. W. (1989). Pooled time series analysis. Thousand Oaks, CA: Sage.

Scarr, S. (1997). Rules of evidence. A larger context for statistical debate. Psychological Science, 8, 16-17.

Schaffer, S. (1989). Glass works: Newton's prisms and the uses of experiment. In D. Gooding, T. Pinch, & S. Schaffer (Eds.), The uses of experiment: Studies in the natural sciences (pp. 105-114). Cambridge, England: Cambridge University Press.

Schaffner, K. F. (Ed.). (1986). Ethical issues in the use of clinical controls. Journal of Medical Philosophy, 11, 297-404.

Scharfstein, D. O., Rotnitiky, A., & Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. Journal of the American Statistical Association, 94, 1096-1120.

Schlesselman, J. J. (1982). Case-control studies: Design, conduct, analysis. New York: Oxford University Press.

Schmidt, D., & Leppik, I. E. (Eds.). (1988). Compliance in epilepsy. Amsterdam: Elsevier.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. Psychological Methods, 1, 115-129.

Schmidt, F. L., & Huntet, J. E. (1977). Development of a general solution to the problem of validity generalization. Journal of Applied Psychology, 62, 529-540.

Schmitt, F. F. (1995). Truth: A primer. Boulder, CO: Westview Press.

Schoenberg, R. (1989). Covariance structure models. Annual Review of Sociology, 15, 425-440.

Schonemann, P. H. (1991). In praise of randomness. Behavioral and Brain Sciences, 14, 162-163.

Schulz, K. F. (1995). Subverting randomization in controlled trials. Journal of the American Medical Association, 274, 1456-1458.

Schulz, K. F., Chalmers, I., Hayes, R. J., & Altman, D. G. (1995). Empirical evidence of bias: Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. Journal of the American Medical Association, 273, 408-412.

Schumacher, J. E., Milby, J. B., Raczynski, J. M., Engle, M., Caldwell, E. S., & Carr, J. A. (1994). Demoralization and threats to validity in Birmingham's Homeless Project. In K. J. Conrad (Ed.), Critically evaluating the role of experiments (pp. 41-44). San Francisco: Jossey-Bass.

Schweinhart, L. J., Barnes, H. V., & Weikart, D. P. (1993). Significant benefits: The High/Scope Perry Preschool study through age 27. (Available from High/Scope Foundation, 600 North River Street, Ypsilanti, MI 48198)

Scientists quibble on calling discovery" planets." (2000, October 6). The Memphis Commercial Appeal, p. A5.

Scott, A. G., & Sechrest, L. (1989). Strength of theory and theory of strength. Evaluation and Program Planning, 12, 329-336.

Scriven, M. (1976). Maximizing the power of causal investigation: The Modus Operandi method. 1n G. V. Glass (Ed.), Evaluation studies review annual (Vol. 1, pp. 101-118). Newbury Park, CA: Sage.

Scriven, M. (1980). The logic of evaluation. Inverness, CA: Edgepress.

Scriven, M. (1991). Evaluation thesaurus (4th ed.). Thousand Oaks, CA: Sage.

Seamon, F., & Feiock, R. C. (1995). Political participation and city/county consolidation: Jacksonville-Duval County. International Journal of Public Administration, 18, 1741-1752.

Seaver, W. B. (1973). Effects of naturally induced teacher expectancies. Journal of Personality and Social Psychology, 28, 333-342.

Seaver, W. B., & Quartan, R. J. (1976). Regression-discontinuity analysis of dean's list effects. Journal of Educational Psychology, 66, 459-465.

Sechrest; L., West, S. G., Phillips, M A., Redner, R., & Yeaton, W. (1979). Some neglected problems in evaluation research: Strength and integrity of research. In L. Sechrest, S. G. West, M.A. Phillips, R. Redner, & W. Yeaton (Eds.), Evaluation studies review annual (Vol. 4, pp. 15-35). Beverly Hills, CA: Sage.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? Psychological Bulletin, 105, 309-316.

Seligman, M. E. P. (1969). Control group and conditioning: A comment on operationalism. Psychological Review, 76, 484-491.

SenGupta, S. (1995). A similarity-based single study approach to construct and external validity. Dissertation Abstracts International; 55(11), 3458A. (University Microfilms No. 9509453)

Serlin, R. A., & Lapsley, D. K. (1993). Rational appraisal of psychological research and the good-enough principle. In G. Keren & C. Lewis (Eds.), A handbook for data analysis in the behavioral sciences: Methodological issues (pp. 199-228). Hillsdale, NJ: Erlbaum.

Shadish, W. R. (1984). Policy research: Lessons from the implementation of deinstitutionalization. American Psychologist, 39, 725-738. ·

Shadish, W. R. (1989). Critical multiplism: A research strategy and its attendant tactics. In L. Sechrest, H. Freeman, & A. Mulley (Eds.), Health services research methodology: A focus on AIDS (DHHS Publication No. PHS 89-3439, pp. 5-28). Rockville, MD: U.S. Department of Health and Human Services, Public Health Service, National Center for Health Services Research and Health Care Technology Assessment.

Shadish, W. R. (1992a). Do family and marital psychotherapies change what people do? A meta-analysis of behavioral outcomes. InT. D. Cook, H. M. Cooper, D. S. Cordray, H. Hartmann, L. V. Hedges, R. J, Light, T. A. Louis, & F. Mosteller (Eds.), Meta-analysis for explanation: A casebook (pp. 129-208). New York: Russell Sage Foundation.

Shadish, W. R. (1992b, August). Mediators and moderators in psychotherapy meta- analysis. Paper presented at the aunual convention of the American Psychological Association, Washington, DC.

Shadish, W. R. (1994). Critical multiplism: A research strategy and its attendant tactics. In L. B. Sechrest & A. J. Figueredo (Eds.), New directions for program evaluation (pp. 13-57). San Francisco: Jossey-Bass.

Shadish, W. R. (1995a). Philosophy of science and the quantitative-qualitative debates: Thirteen conunon errors. Evaluation and Program Planning, 18, 63-75.

Shadish, W. R. (1995b). The logic of generalization: Five principles common to experiments and ethnographies. American Journal of Community Psychology, 23, 419-428.

Shadish, W. R. (1996). Meta-analysis and the exploration of causal mediating processes: A primer of examples, methods, and issues. Psychological Methods, 1, 47-65:

Shadish, W. R. (2000). The empirical program of quasi-experimentation. In L. Bickman (Ed.), Validity and social experimentation: Donald Campbell's legacy (pp. 13-35). Thousand Oaks, CA: Sage.

Shadish, W. R., & Cook, T. D. (1999). Design rules: More steps towards a complete theory of quasi-experimentation. Statistical Science, 14, 294-300.

Shadish, W. R., Cook, T. D., & Houts, A. C. (1986). Quasi-experimentation in a critical multiplist mode. In W. M. K. Trochim (Ed.), Advances in quasi-experimental design and analysis (pp. 29-46). San Francisco: Jossey-Bass.

Shadish, W. R, Cook, T. D., & Leviton, L. C. (1991). Foundations of program evaluation: Theories of practice. Newbury Park, CA: Sage.

Shadish, W. R., Doherty, M., & Montgomery, L. M. (1989). How many studies are in the file drawer? An estimate from the family/marital psychotherapy literature. Clinical Psychology Review, 9, 589-603.

Shadish, W. R., & Fuller, S. (Eds.). (1994). The social psychology of science. New York: Guilford Press.

Shadish, W. Rl, Fuller, S., Gorman, M. E., Amabile, T. M., Kruglanski, A. W., Rosenthal, R., & Rosenwein, R. E. (1994). Social psychplogy of science: Aconceptual and research program. In W. R. Shadish & S. Fuller (Eds.), Social psychology of science (pp. 3-123). New York: Guilford Press.

Shadish, W. R., & Haddock, C. K. (1994). Combining estimates of effect size. In H. M. Cooper & L. V. Hedges (Eds.), The handbook of research synthesis (pp. 261-281). New York: Russell Sage Foundation.

Shadish, W. R., & Heinsman, D. T. (1997). Experiments versus quasi-experiments: Do you get the same answer? In W. J. Bukoski (Ed.), Meta-analysis of drug abuse prevention programs (NIDA Research Monograph, DHHS Publication No. ADM 97-170, pp. 147-164). Washington, DC: Superintendent of Documents.

Shadish, W. R., Hu, X., Glaser, R. R., Kownacki, R. J., & Wong, T. (1998). A method for exploring the effects of attrition in randomized experiments with dichotomous outcomes. Psychological Methods, 3, 3-22.

Shadish, W. R., Matt, G. E., Navarro, A. M., & Phillips, G. (2000). The effects of psychological therapies under clinically representative conditions: A meta-analysis, Psychological Bulletin, 126, 512-529.

Shadish, W. R., Matt, G., Navarro, A., Siegle, G., Crits-Christoph, P., Hazelrigg, M., Jorm, A., Lyons, L. S., Nietzel, M. T., Prout, H. T., Robinson, L., Smith, M. L., Svartberg, M., & Weiss, B. (1997). Evidence that therapy works in clinically representative conditions. Journal of Consulting and Clinidal Psychology, 65, 355-365.

Shadish, W. R., Montgomery, L. M., Wilson, P., Wilson, M. R., Bright, I., & Okwumabua, T. (1993). The effects of family and marital psychotherapies: A meta-analysis. Jornal of Consulting and Clinical Psychology, 61, 992-1002.

Shadish, W. R., Newman, D. L., Scheirer, M. A., & Wye, C. (1995). Developing the guiding principles. In W. R. Shadish, D. L. Newman, M. A. Scheirer, & C. Wye (Eds.), Guiding principles for evaluatord (pp. 3-18). San Francisco: Jossey-Bass.

Shadish, W. R., &. Ragsdale, K. (1996). Random versus nonrandom assignment in psychotherapy experiments: rip you get the same answer? Journal of Consulting and Clinical Psychology, 64, 1290-1305.

Shadish, W. R., & Reis, J. (1984). A review of studies of the effectiveness of programs to improve pregnancy outcome. Evaluation Review, 8, 747-776.

Shadish, W. R., Robinson, L., & Lu, C, (1999). ES: A computer program and manual for effect size calculation. St. Paul, MN: Assessment Systems Corporation.

Shadish, W. R., Silber, B., Orwin, R. G., & Bootzin, R. R. (1985). The subjective well- being of mental patients in nursing homes. Evaluation and Program Planning, 8, 239-250.

Shadish, W. R., Straw, R. B., McSweeny, A. J., Koller, D. L., & Bootzin, R. R.(1981). Nursing home care for mental patients: Descriptive data and some propositions. American Journal of Community Psychology, 9, 617-633.

Shadish, W. R., & Sweeney, R. (1991). Mediators and moderators in meta-analysis: There's a reason we don't let dodo pirds tell us which psychotherapies should have prizes. Journal of Consulting and Clinical Psychology, 59, 883-893.

Shapin, S. (1994). A social history of truth: Civility and science in seventeenth-century England. Chicago: University of Chicago Press.

Shapiro, A. K., & Shapiro, E. (1997). The powerful placebo. Baltimore: Johns Hopkins University Press.

Shapiro, J. Z. (1984). The social costs of methodological rigor: A note on the problem of massive attrition. Evaluation Review, 8, 705-712.

Sharpe, T. R., & Wetherbee, H. (1980). Final report: Evaluation of the Improved Pregnancy Outcome Program. Tupelo, MS: Mississippi State Board of Health, Three Rivers District Health Department.

Shaw, R. A., Rosati, M. J., Salzman, P., Coles, C. R., & McGeary, C. (1997). Effects on adolescent ATOD behaviors and attitudes of a 5-year community partnership. Evaluation and Program Planning, 20, 307-313.

Sherif, J., Harvey, O. J., White, B. J., Hood, W. R., & Sherif, C. W. (1961). Intergroup conflict and cooperation: The Robbers Cave Experiment. Norman: University of Oklahoma Book Exchange.

Sherman, L. W., & Berk, R. A. (1985). The randomization of arrest. In R. F. Boruch & W. Wothke (Eds.), Randomization and field experimentation (pp. 15-25). San Francisco: Jossey-Bass.

Shih, W. J., & Quan, H. (1997). Testing for treatment differences with dropouts present in clinical trials: A composite approach. Statistics in Medicine, 16, 1225-1239.

Shoham-Salomon, V., & Rosenthal, R. (1987). Paradoxical interventions: A meta-analysis. Journal of Consulting and Clinical Psychology, 55, 22-28.

ShOnemann, P. H. (1991). In praise of randomness. Behavioral and Brain Sciences, 14, 162-163.

Shonkoff, J. P., & Phillips, D. A. (Eds.). (2000). From neurons to neighborhoods: The science of early childhood development. Washington, DC: National Academy Press.

Shrout, P. E. (1997). Should significance tests be banned? Introduction to a special section exploring the pros and cons. Psychological Science, 8, 1-2.

Shumaker, S. A., & Rejeski, W. J, (Eds.). (2000). Adherence to behavioral and pharmacological interventions in clinical research in older adults [Special issue]. Controlled Clinical Trials, 21(5S).

Shumway, R. H. (1988). Applied statistical time series analysis. Englewood Cliffs, NJ: Prentice-Hall.

Sidman, M. (1960). Tactics of scientific research: Evaluating experimental data in psychology. Boston: Authors Cooperative.

Sieber, J. E. (1992). Planning ethically responsible research: A guide for students and internal review boards. Newbury Park, CA: Sage.

Siegel, A. E., & Siegel, S. (1957). Reference groups, membership groups, and attitude change. Journal of Abnormal and Social Psychology, 55, 360-364.

Siemiatycki, J. (1989). Friendly control bias. Journal of Clinical Epidemiology, 42, 687-688.

Siemiatycki, J., Colle, S., Campbell, S., Dewar, R., & Belmonte, M. M. (1989). Case-control study of insulin dependent (type I) diabetes mellitus. Diabetes Care, 12, 209-216.

Silka, L, (1989). Intuitive judgments of change. New York: Springer-Verlag.

Silliman, N. P. (1997). Hierarchical selection models with applications in meta-analysis. Journal of the American Statistical Association, 92, 926-936.

Silverman, W. A. (1977). The lesson of retrolental fibroplasia. Scientific American, 236, 100-107.

Simes, R. J. (1987). Confronting publication bias: A cohort design for meta-analysis. Statistics in Medicine, 6, 11-29.

Simon, H. A. (1976). Administrative behavior. New York: Free Press.

Simpson, J. M., Klar, N., & Donner, A. (1995). Accounting for cluster randomization: A review of primary prevention trials, 1990 through 1993. American journal of Public Health, 85, 1378-1383.

Skinner, B. F. (1961). Cumulative record. New York: Appleton-Century-Crofts.

Slavin, R. E. (1986). Best-evidence synthesis: An alternative to meta-analysis and traditional reviews. Educational Researchers, 15, 5-11.

Smith, B., & Sechrest, L. (1991). Treatment of aptitude x treatment interactions. Journal of Consulting and Clinical Psychology, 59, 233-244.

Smith, E. E., & Medin, D. L. (1981). Categories and concepts. Cambridge, MA: Harvard University Press.

Smith, E. R. (1992). Beliefs, attributions, and evaluations: Nonhierarchical models of mediation in social cognition. Journal of Personality and Social Psychology, 43, 248-259.

Smith, G. (1997). Do statistics test scores regress toward the mean? Chance, 10, 42-45.

Smith, H. (1997). Matching with multiple controls to estimate treatment effects in observational studies. Sociological Methodology, 27, 325-353.

Smith, M. L. (1980). Publication bias and meta-analysis. Evaluation in Education, 4, 22-24.

Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. American Psychologist, 32, 752-760.

Smith, M. L., Glass, G. V., & Miller, T.l. (1980). The benefits of psychotherapy. Baltimore: Johns Hopkins University Press.

Smith, N. L. (Ed.). (1992). Varieties of investigative journalism. San Francisco: Jossey-Bass.

Smoot, S. L. (1989). Meta-analysis of single subject research in education: A comparison of four metrics (Doctoral dissertation, Georgia State University). Dissertation Abstracts International, 50(07), 1928A.

Snow, R. E. (1991). Aptitude-treatment interactions as a framework for research on individual differences in psychotherapy. Journal of Consulting and Clinical Psychology, 59, 205-216.

Snowdon, C., Elbourne, D., & Garcia, J. (1999). Zelen randomization: Attitudes of parents participating in a neonatal clinical trial. Controlled Clinical Trials, 20, 149-171.

Snijders, T. A. B., & Bosker, R. J. (1999). Multilevel analysis: An introduction to basic and advanced multilevel modeling. London: Sage.

Snyder, D. K., & Wills, R. M. (1989). Behavioral versus insight-oriented marital therapy: Effects on individual and interspousal functioning. Journal of Consulting and Clinical Psychology, 57, 39-46.

Snyder, D. K., Wills, R. M., & Grady-Fletcher, A. (1991). Long-term effectiveness of behavioral versus insight-oriented marital therapy: A 4-year follow-up study. Journal of Consulting and Clinical Psychology, 59, 138-141.

Sobel, M E. (1993). Causal inference in the social and behavioral sciences. In G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), Handbook for statistical modeling in the social and behavioral sciences (pp. 1-38). New York: Plenum.

Sobel, M. E. (2000). Causal inference in the social sciences. Journal of the American Statistical Association, 95, 647-650.

Solomon, R. L. (1949). An extension of control group design. Psychological Bulletin, 46, 137-150.

Sommer, A., & Zeger, S. L. (1991). On estimating efficacy from clinical trials. Statistics in Medicine, 10, 45-52.

Sommer, B. (1987). The file drawer effect and publication rates in menstrual cycle research. Psychology of Women Quarterly, 11, 233-242.

Sorensen, G., Emmons, K., Hunt, M., & Johnston, D. (1998). Implications of the results of community trials. Annual Review of Public Health, 19, 379-416.

Speer, D. C. (1994). Can treatment research inform decision makers? Nonexperimental method issues and examples among older outpatients. Journal of Consulting and Clinical Psychology, 62, 560-568.

Speer, D. C., & Swindle, R. (1982). The '"monitoring model" and the mortality x treatment interaction threat to validity in mental health outcome evaluation. American Journal of Community Psychology, 10, 541-552.

Spiegelhalter, D. J., Freedman, L. S., & Blackburn, P.R. (1986). Monitoring clinical trials: Conditional or predictive power? Controlled Clinical Trials, 7, 8-17.

Spirtes, P., Glymour, C., & Scheines, R. (2000). Causation, prediction and search (2nd ed.). Cambridge, MA: MlT Press.

Spirtes, P., Scheines, R., & Glymour, C. (1990a). Reply to comments. Sociological Methods and Research, 19, 107-121.

Spirtes, P., Scheines, R., & Glymour, C. (1990b). Simulation studies of the reliability of computeraided model specification using the TETRAD II, EQS, and LISREL programs. Sociological Methods and Research, 19, 3-66.

Spitz, H. H. (1993). Were children randomly assigned in the Perry Preschool Project? American Psychologist, 48, 915.

Stadthaus, A. M. {1972). A comparison of the subsequent academic achievement of marginal selectees and rejectees for the Cincinnati public schools special college preparatory program: An application of

Campbell's regression discontinuity design (Doctoral dissertation, University of Cincinnati, 1972). Dissertation Abstracts International, 33{06), 2825A.

Staines, G. L., McKendrick, K., Perlis, T., Sacks, S., & DeLeon, G. {1999). Sequential assignment and treatment-as-usual: Alternatives to standard experimental designs in field studies of treatment efficacy. Evaluation Review, 23, 47-76.

Stake, R. E., & Trumbull, D. J. (1982). Naturalistic generalizations. Review Journal of Philosophy and Social Science, 7, 1-12.

Stanley, B., & Sieber, J. E. (Eds.). (1992). Social research on children and adolescents: Ethical issues. Newbury Park, CA: Sage.

Stanley, T. D. (1991). "Regression-discontinuity design" by any other name might be less problematic. Evaluation Review, 15, 605-624.

Stanley, T. D., & Robinson, A. (1990). Sifting statistical significance from the artifact of regression-discontinuity design. Evaluation Review, 14, 166-181.

Stanley, W. D. (1987). Economic migrants or refugees from violence? A time series analysis of Salvadoran migration to the United States. Latin American Law Review, 12, 132-154.

Stanton, M.D., & Shadish, W. R. (1997). Outcome, attrition and family-couples treatment for drug abuse: A meta-analysis and review of the controlled, comparative studies. Psychological Bulletin, 122, 170-191.

Starfield, B. (1977). Efficacy and effectiveness of primary medical care for children. In Harvard Child Health Project, Children's medical care needs and treatment: Report of the Harvard Child Health Project. Cambridge, MA: Ballinger.

Statistical Solutions. (1998). SOLAS for Missing Data Analysis 1.0 [Computer software]. (Available from Statistical Solutions, 8 South Bank, Crosse's Green, Cork, Ireland)

Steering Committee of the Physicians' Health Study Research Group. (1988). Preliminary report: Findings from the aspirin component of the ongoing physicians' health study. New England Journal of Medicine, 318, 271-295.

Stein, M. A., & Test, L.I. (1980).Alternative to mental hospital treannent: I. Conceptual model, treatment program, clinical evaluation. Archives of General Psychiatry, 37, 392-397.

Steiner, M. S., & Gingrich, J. R. (2000). Gene therapy for prostate cancer: Where are we now? Journal of Urology, 164, 1121-1136.

Stelzl, I. (1986). Changing a causal hypothesis without changing the fit: Some rules for generating equivalent path models. Multivariate Behavioral Research, 21, 309-331.

Stevens, S. J. (1994). Common implementation issues in three large-scale social experiments. In K. J. Conrad (Ed.), Critically evaluating the role of experiments (pp. 45-53). San Francisco: Jossey-Bass.

Stewart, A. L., Sherbourne, C. D., Wells, K. B., Burnam, M.A., Rogers, W. H., Hays, R. D., & Ware, J. E. (1993). Do depressed patients in different treannent settings have different levels of well-being and functioning? Journal of Consulting and Clinical Psychology, 61, 849-857.

Stigler, S. M. (1986). The history of statistics: The measurement of uncertainty before 1900. Cambridge, MA, and London, England: Harvard University/Belnap Press.

Stock, W. A. (1994). Systematic coding for research synthesis. In H. Cooper & L. V. Hedges (Eds.), The handbook of research synthesis (pp. 126-138). New York: Russell Sage Foundation.

Stolzenberg, R., & Relies, D. (1990). Theory testing in a world of constrained research design: The significance of Heckman's censored sampling bias correction for nonexperimental research. Sociological Methods and Research, 18, 395-415.

Stone, R. A., Obrosky, D. S., Singer, D. E., Kapoor, W. N., Fine, M. J., & The Pneumonia Patient Outcomes Research Team (PORT) Investigators. (1995). Propensity score adjustment for pretreatment differences between hospitalized and ambulatory patients with community-acquired pneumonia. Medical Care, 33 (Suppl.), AS56-AS66.

Stout, R. L., Brown, P. J., Longabaugh, R., & Noel, N. (1996). Determinants of research follow-up participation in an alcohol treatment outcome trial. Journal of Consulting and Clinical Psychology, 64, 614-618.

Stromsdorfer, E. W., & Farkas, G. (Eds.). (1980). Evaluation studies review annual (Vol. 51. Beverly Hills, CA: age.

Stroup, D. F., Berlin, J. A., Morton, S.C., Olkin, I., Williamson, G. D., Rennie, D., Moher, D., Becker, B. J., Sipe, T. A., Thacker, S. B., for the Meta-Analysis of Observational Studies in Epidemiology (MOOSE) Group. (2000). Meta-analysis of observational studies in epidemiology: A proposal for reporting. Journal of the American Medical Association, 283, 2008-2012.

Sullivan, C. M., Rumptz, M. H., Campbell, R., Eby, K. K., & Davidson, W. S. (1996). Retaining participants in longitudinal community research: A comprehensive protocol. Journal of Applied Behavioral Science, 32, 262-276.

Swanson, H. L., & Sqchselee, C. (2000). A meta-analysis of single-subject-design intervention research for students with LD. Journal of Learning Disabilities, 33, 114-136.

Tallinadge, G. K., & Horst, D.P. (1976). A procedural guide for validating achievement gains in educational projects. (Evaluation in Education Monograph No. 2). Washington, DC: U.S. Department of Health, Education, and Welfare.

Tallmadge, G. K., & Wood, C. T. (1978). User's guide: ESEA Title l evaluation and reporting system. Mountain View, CA: RMC Research.

Tamura, R.N., Faries, D. E., Andersen, J. S., & Heiligenstein, J. H. (1994). A case study of an adaptive clinical trial in the treatment of out-patients with depressive disorder. Journal of the American Statistical Association, 89, 768-776.

Tan, M., & Xiong, X. (1996). Continuous and group sequential conditional probability ratio tests for Phase II clinical trials. Statistics in Medicine, 15, 2037-2051.

Tanaka, J. S., & Huba, G. J. (1984). Confirmatory hierarchical factor analysis of psychological distress measures. Journal of Personality and Social Psychology, 46, 621-635.

Tanaka, J. S., Panter, A. T., Winborne, W. C., & Huba, G. J. (1990). Theory testing in personality and social psychology with structural equation models: A primer in 20 questions. In C. Hendrick & M. S. Clark (Eds.), Research methods in personality and social psychology (pp. 217-242). Newbury Park, CA: Sage.

Taylor, S. J., & Bogdan, R. (1984). Introduction to qualitative research methods: The search for meanings. New York: Wiley.

TayloR, G. J., Rubin, R., Tucker, M., Greene, H. L., Rudikoff, M. D., & Weisfeldt, M L. (1978). External cardiac .compression: A randomized comparison of mechanical and manual techniques. Journal of the American Medical Association, 240, 644-646.

Teague, M. L., Bernardo, D. J., & Mapp, H. P. (1995). Farm-level economic analysis incorporating stochastic environmental risk assessment. American Journal of Agricultural Economics, 77, 8-19.

Test, M.A., & Burke, S. S. (1985). Random assignment of chronically mentally ill persons to hospital or community treatment. In R. F. Boruch & W. Wothke (Eds.), Randomization and field experimentation (pp. 81-93), San Francisco: jossey-Bass.

Test, M.A., & Stein, L. I. (1980). Alternative to mental hospital treatment: III. Social cost. Archives of Geneal Psychiatry, 37, 409-412.

Tester, K. (1993). The life and times of post-modernity. London: Routledge.

Thistlewaite, D. L., & Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. Journal of Educational Psychology, 51, 309-317.

Thomas, A., Spiegelhalter, D. J., & Gilks, W. R. (1992). BUGS: A program to perform Bayesian inference using Gibbs sampling. In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), Bayesian statistics 4. Oxford, England: Clarendon Press.

Thomas, G. B. (1997). A program evaluation of the remedial and developmental studies program at Tennessee State University (Doctoral dissertation, Vanderbilt University, 1997). Dissertation Abstracts International, 58(08), 3042A.

Thomas, L., & Krebs, C. J. (1997). A review of statistical power analysis software. Bulletin of the Ecological Society of America, 78, 128-139.

Thompson, B. (1993). Statistical significance testing in contemporary practice: Some proposed alternatives with comments from journal editors [special issue]. Journal of Experimental Education, 61(4).

Thompson, D. C., Rivara, F. P., & Thompson, R. (2000). Helmets for preventing head and facial injuries in bicyclists (Cochrane Review). The Cochrane Library, Issue 3. Oxford, England: Update Software.

Thompson, S. K. (1992). Sampling. New York: Wiley.

Tilden, V. P., & Shepherd, P. (1987), Increasing the rate of identification of battered women in an emergency department: Use of a nursing protocol. Research in Nursing and Health, 10,209-215.

Time series database of U.S. and international statistics ready for manipulation. (1992). Database Searcher, 8, 27-29.

Tinbergen, J. (1956). Economic policy principles and design. Amsterdam: North-Holland.

Tong, H. (1990). Non-linear time series: A dynamical system approach. New York: Oxford University Press.

Toulmin, S. E. (1961). Foresight and understanding: An inquiry into the aims of science. Bloomington: Indiana University Press.

Tracey, T. J., Sherry, P., & Keitel, M. (1986). Distress and help-seeking as a function of person-erivironment fit and self-efficacy: A causal model. American Journal of Community Psychology, 14, 657-676.

Trend, M.G. (1979). On the reconciliation of qualitative and quantitative analyses: A case study. In T. D. Cook & C. S. Reichardt (Eds.), Qualitative and quantitative methods in evaluation research (pp. 68-86). Newbury Park, CA: Sage.

Triplett, N. (1898). The dynamogenic factors in pacemaking and competition. American Journal of Psychology, 9, 507-533.

Trochim, W. M. K. (1980). The regression-discontinuity design in Title I evaluation: Implementation, analysis, and variations. Unpublished doctoral dissertation, Northwestern University, Evanston, IL.

Trochim, W. M. K. (1984). Research design for program evaluation: The regression-discontinuity approach. Newbury Park, CA: Sage.

Trochim, W. M. K. (1985). Pattern matching, validity, and conceptualization in program evaluation. Evaluation Review, 9, 575-604.

Trochim, W. M. K. (1990). The regression discontinuity design. In L. Sechrest, E. Perrin, & J. Bunker (Eds.), Research methodology: Strengthening causal interpretations of nonexperimental data (pp. 119-140). Rockville, MD: Public Health Service, Agency for Health Care Policy and Research.

Trochim, W. M. K., & Campbell, D. T. (1996). The regression point displacement design for evaluating community-based pilot programs and demonstration projects. Unpublished manuscript. (Available from the Department of Policy Analysis and Management, Cornell University, Room N136C, MVR Hall, lthaca, NY 14853)

Trochim, W. M. K., & Cappelleri, J. C. (1992). Cutoff assignment strategies for enhancing randomized clinical trials. Controlled Clinical Trials, 13, 190-212.

Trochim, W. M. K., Cappelleri, J. C., & Reichardt, C. S. (1991). Random measurement error does not bias the treatment effect estimate in the regression-discontinuity design: II. When an interaction effect is present. Evaluation Review, 15, 571-604.

Tufte, E. R. (1983). The visual display of quantitative information. Cheshire, CT: Graph- ics Press.

Tufte, E. R. (1990). Envisioning information. Cheshire, CT: Graphics Press.

Tukey, J. W. (1991). The philosophy of multiple comparisons. Statistical Science, 6, 100-116.

Turpin, R. S., & Sinacore, J. M. (Eds.). (1991). Multisite evaluations. San Francisco: Jossey-Bass.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. Science, 185, 1124-1131.

Uebel, T. E. (1992). Overcoming logical positivism from within: The emergence of Neurath's naturalism in the Vienna Circle's protocol sentence debate. Amsterdam and Atlanta, GA: Editions Rodopi B.V.

Valins, S., & Baum, A. (1973). Residential group size, social interaction, and crowding. Environment and Behavior, 5, 421-439.

Varnell, S., Murray, D. M., & Baker, W. L. (in press). An evaluation of analysis options for the one group per condition design: Can any of the alternatives overcome the problems inherent in this design? Evaluation Review.

Veatch, R., & Sellitto, S. (1973). Human experimentation: The ethical questions persist. Hastings Center Report, 3, 1-3.

Velicer, W. F. (1994). Time series models of individual substance abusers. In L. M Collins & L. A. Seitz (Eds.), Advances in data analysis for prevention intervention research (NIDA Research Monograph No. 142, pp. 264-299). Rockville MD: National Institute on Drug Abuse.

Velicer, W. F., & Harrop, J. (1983). The reliability and accuracy of time series model identification. Evaluation Review, 7, 551-560.

Veney, J. E. (1993). Evaluation applications of regression analysis with time-series data. Evaluation Practice, 14, 259-274.

Verbeek, M., & Nijman, T. (1992). Testing for selectivity bias in panel data models. International Economic Review, 33, 681-703.

Vinokur, A. D., Price, R. H., & Caplan, R. D. (1991). From field experiments to program implementation: Assessing the potential outcomes of an experimental intervention program for unemployed persons. American Journal of Community Psychology, 19, 543-562.

Vinovskis, M.A. (1998). Changing federal strategies for supporting educational research, development and statistics. Unpublished manuscript, National Educational Research Policy and Priorities Board, U.S. Department of Education.

Virdin, L. M. (1993). A test ofthe robustness of estimators that model selection in the nonequivalent control group design. Unpublished doctoral dissertation, Arizona State University, Tenipe.

Vessey, M.P. (1979). Comment. Journal of Chronic Diseases, 32, 64-66.

Visser, R. A., & deLeeuw, J. (1984). Maximum likelihood analysis for a generalized regression-discontinuity design. Journal of Educational Statistics, 9, 45-60.

Viswesvaran, C., & Schmidt, F. L. (1992). A meta-analytic comparison of the effectiveness of smoking cessation methods. Journal of Applied Psychology, 77, 554-561.

Vosniadou, S., & Ortony, A. (1989). Similarity and analogical reasoning: A synthesis. In S. Vosniadou & A. Ortony (Eds.), Similarity and analogical reasoning (pp. 1-17). New York: Cambridge University Press.

WagenaaJ; A. C., & Holder, H. D. (1991). A change from public to private sale of wine: Results from natural experiments in Iowa and West Virginia. Journal of Studies on Alcohol, 52, 162-173.

Wagoner, J. L. (1992). The contribution of group therapy to the successful completion of probation for adult substance abusers. Dissertation Abstracts International, 53(03), 724A. (University Microfilms No. AAC92-20873)

Wainer, H. (Ed.). (1986). Drawing inferences from self-selected samples. New York: Springer-Verlag.

Wallace, L. W. (1987). The Community Penalties Act of 1983: An evaluation of the law, its implementation, and its impact in North Carolina. Unpqblished doctoral dissertation, University of Nebraska.

Wallace, P., Cutler, S., & Haines, A. (1988). Randomised controlled trial of general practitioner intervention in patients with excessive alcohol consumption. British Medical Journal, 297, 663-668.

Walther, B. J., & Ross, A. S. (1982). The effect on behavior of being in a control group. Basic and Applied Social Psychology, 3, 259-266.

Wampler, K. S., & Serovich, J. M. (1996). Meta-analysis in family therapy research. In D. H. Sprenkle & S. M. Moon (Eds.), Research methods in family therapy (pp. 286-303). New York: Guilford Press.

Wampold, B. E. (1992). The intensive examination of social interactions. In T. R. Kratochwill & J. R. Levin (Eds.), Single-case research design and analysis (pp. 93-131). Hillsdale, NJ: Erlbaurn.

Wampold, B. E., & Furlong, M. J. (1981). The heuristics of visual inspection. Behavioral Assessment, 3, 79-92.

Wampold, B. E., Mondin, G. W., Moody, M., Stich, F., Benson, K., & Ahn, H. (1997). A meta-analysis of outcome studies comparing bona fide psychotherapies: Empirically, "All must have prizes." Psycholosical Bulletin, 122, 203-215.

Wampold, B. E., & Worsham, N. L. (1986). Randomization tests for multiple-baseline designs. Behavioral Assessment, 8, 135-143.

Wang, M. C., & Bushman, B. J. (1998). Using the normal quantile plot to explore meta-analytic data sets. Psychological Methods, 3, 46-54.

Wang, M. C., & Bushman, B. J. (1999). Integrating results through meta-analytic review using SAS software. Cary, NC: SAS Institute.

Washington v. Davis, 426 U.S. 229 (1976).

Watts, H., & Rees, A.W. (1976). The New Jersey income-maintenance experiment: Vol. 2. Labor-supply responses. New York: Academic Press.

Webb, E. J., Campbell, D. T., Schwartz, R. D., & Sechrest, L. (1966). Unobtrusive measures. Skokie, IL: Rand McNally.

Webb, E. J., Campbell, D. T., Schwartz, R. D., Sechrest, L., & Grove, J.B. (1981). Nonreactive measures in the social sciences (2nd ed.). Boston, MA: Houghton Mifflin.

Webb, J. F., Khazen, R. S., Hanley, W. B., Partington, M. S., Percy, W. J. L., & Rathburn, J. C. (1973). PKU screening: Is it worth it? Canadian Medical Association Journal, 108, 328-329.

Weber, S. J., Cook, T. D., & Campbell, D. T. (1971). The effects of school integration on the academic self-concept of public-school children. Paper presented at the annual meeting of the Midwestern Psychological Association, Detroit, MI.

Wei, L. J. (1978). An application of an urn model to the design of sequential controlled trials. Journal of the American Statistical Association, 73, 559-563.

Wei, W. W. S. (1990). Time series analysis: Univariate and multivariate methods. Redwood City, CA: Addison-Wesley.

Weiss, B., Williams, J. H., Margen, S., Abrams, B., Caan, B., Citron, L. j., Cox, C., McKibben, J., Ogar, D., & Schultz, S. (1980). Behavioral responses to artificial food colors. Science, 207, 1487-1489.

Weiss, C. H. (1980). Knowledge creep and decision accretion. Knowledge: Creation, Diffusion, Utilization, 1, 381-404.

Weiss, C. H. (1988). Evaluation for decisions: Is anybody there? Does anybody care? Evaluation Practice, 9, 5-20.

Weiss, C. H. (1998). Evaluation (2nd ed.). Upper Saddle River, NJ: Prentice-Hall.

Weiss, C. H., & Bucuvalas, M. J. (1980). Social science research and decision-making. New York: Columbia University Press.

Weisz, J. R., Weiss, B., & Donenberg, G. R. (1992). The lab versus the clinic: Effects of child and adolescent psychotherapy. American Psychologist, 47, 1578-1585.

Weisz, J. R., Weiss, B., & Langmeyer, D. B. (1987). Giving up on child psychotherapy: Who drops out? Journal of Consulting and Clinical Psychology, 55, 916-918.

Welch, W. P., Frank, R. G., & Costello, A. J. (1983). Missing data in psychiatric research: A solution. Psychological Bulletin, 94,177-180.

West, S. G., Finch, J. F., & Curran, P. J. (1995). Structural equation models with nonnormal variables: Problems and remedies. In R. Hoyle (Ed.), Structural equation modeling: Issues and applications (pp. 56-75). Thousand Oaks, CA: Sage.

West, S. G., & Hepworth, J. T. (1991). Statistical issues in the study of temporal data: Daily experiences. Journal of Personality, 59, 609-662.

West, S. G., Hepworth, J. T., McCall, M. A., & Reich, J. W. (1989). An evaluation of Arizona's July 1982 drunk driving law: Effects on the City of Phoenix. Journal of Applied Social Psychology, 19, 1212-1237.

West, S. G., & Sagarin, B. (2000). Subject selection and loss in randomized experiments. In L. Bickman (Ed.), Research design: Donald Campell's legacy (Vol. 2, pp. 117-154). Thousand Oaks, CA: Sage.

Westlake, W. J. (1988). Bioavailability and bioequivalence of pharmaceutical formulations. In K. E. Peace (Ed.), Biopharmaceutical statistics for drug development (pp. 329-352). New York: Dekker.

Westmeyer, H. (in press). Explanation in the social sciences. In N. J. Smelser & P. B. Baltes (Eds.), International encyclopedia of the social and behavioral sciences. Oxford, England: Elsevier.

White, D. M., Rusch, F. R., Kazdin, A. E., & Hartmann, D. P. (1989). Applications of meta-analysis in individual-subject research. Behavioral Assessment, 11, 281-296.

White, H. (1981). Consequences and detection of misspecified nonlinear regression models. Journal of the American Statistical Association, 76, 419-433.

White, H. D. (1994). Scientific communication and literature retrieval. In H. Cooper & L. V. Hedges (Eds.), The handbook of research synthesis (pp. 41-55). New York: Russell Sage Foundation.

White, L., Tursky, B., & Schwartz, G. E. (Eds.). (1985). Placebo: Theory, research, and mechanisms. New York: Guilford Press.

White, P. A. (1990). Ideas about causation in philosophy and psychology. Psychological Bulletin, 108, 3-18.

Whittier, J.G. (1989). The works of John Greenleaf Whittier (Vol. 1). New York: Houghton Mifflin.

Widom, C. S., Weiler, B. L., & Cottler, L. B. (1999). Childhood victimization and drug abuse: A comparison of prospective and retrospective findings. Journal of Consulting and Clinical Psychology, 67, 867-880.

Wilcox, R. R. (1995). ANOVA: A paradigm for low power and misleading measures of effect size? Review of Educational Research, 65, 51-77.

Wilcox, R. R. (1996). Statistics for the social sciences. New York: Academic Press.

Wilder, C. S. (1972, July). Physician visits, volume, and interval since last visit, U.S., 1969 (Series 10, No. 75; DHEW Pub. No. HSM 72-1064). Rockville, MD: National Center for Health Statistics.

Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. American Psychologist, 54, 594-604.

Willett, J. B. (1988). Questions and answers in the measurement of change. Review of Research in Education, 15, 345-422.

Willett, J. B., & Singer, J. D. (1991). From whether to when: New methods for studying student dropout and teacher attrition. Review of Educational Research, 61, 407-450.

Williams, S. V. (1990). Regression-discontinuity design in health evaluation. In L. Sechrest, B. Perrin, & J. Bunker (Eds.), Health services research methodology: Strengthening causal interpretations of nonexperimental data (DHHS Publication No. PHS 90-3454, pp. 145-149). Rockville, MD: U.S. Department of Health and Human Services.

Willms, J. D. (1992). Monitoring school performance: A guide for educators. Washington, DC: The Falmer Press.

Willner, P. (1991). Methods for assessing the validity of animal models of human psychopathology. In A. A. Boulton, G. B. Baker, & M. T. Martin-Iverson (eds.), Neuromethods (Vol. 18, pp. 1-23). Clifton, NJ: Humana Press.

Willson, V. L., & Putnam, R. R. (1982). A meta-analysis of pretest sensitization effects in experimental design. American Educational Research Journal, 19, 249-258.

Wilson, E. B. (1952). An introduction to scientific research. New York: McGraw-Hill.

Wilson, M. C., Hayward, R. S. A., Tunis, S. R., Bass, E. B. & Guyatt, G. (1995). Users' guides to the medical literature: Part 8. How to use clinical practice guidelines. B. What are the recommendations and will they help you in caring for your patients? Journal of the American Medical Association, 274, 1630-1632.

Winer, B. J., Brown, D. R., & Michels, K. M. (1991). Statistical principles in experimental design (3rd ed.). New York: McGraw-Hill.

Winship, C., & Mare, R. D. (1992). Models for sample selection bias. Annual Review of Sociology, 18, 327-350.

Winship, C., & Morgan, S. L. (1999). The estimation of causal effects from observational data. Annual Review of Sociology, 25, 659-707.

Winston, A. S. (1990). Robert Sessions Woodworth and the "Columbia Bible": How the psychological experiment was redefined. American Journal of Psychology, 103, 391-401.

Winston, A. S., & Blais, D. J. (1996). What counts as an experiment? A transdisciplinary analysis oftextbooks, 1930-1970. American Journal of Psychology, 109, 599-616.

Winston, P. H., & Horn, B. K. P. (1989). LISP (3rd ed.). Reading, MA: Addison-Wesley.

Witte, J. F. (1998). The Milwaukee voucher experiment. Educational Evaluation and Policy Analysis, 20, 229-251.

Wolf, F. M. (1986). Meta-analysis: Quantitative methods for research synthesis. Thousand Oaks, CA: Sage.

Wood, G. (1978). The knew-it-all-along effect. Journal of Experimental Psychology: Human Perception and Performance, 4, 345-353.

Woodworth, G. (1994). Managing meta-analytic data bases. In H. Cooper & L. V. Hedges (Eds.), The handbook of research synthesis (pp. 177-189). New York: Russell Sage Foundation.

World Medical Association. (2000). Declaration of Helsink: Ethical principles for medical research involving human subjects. Journal of the American Medical Association, 284, 3043-3045.

Wortman, C. B., & Rabinowitz, V. C. (1979). Random assignment: The fairest of them all. In L. Sechrest, S. G. West, M. A. Phillips, R. Redner, & W. Yeaton (Eds.), Evaluation studies review annual (Vol. 4, pp. 177-184). Beverly Hills, CA: Sage.

Wortman, P. M. (1992). Lessons from the meta-analysis of quasi-experiments. In F. B. Bryant, J. Edwards, R. S. Tindale, E. J. Posavac, L. Heath, E. Henderson, & Y. Suarez-Balcazar (Eds.), Methodological issues in applied social psychology (pp. 65-81). New York: Plenum Press.

Wortman, P.M. (1994). Judging research quality. In H. Cooper & L. V. Hedges (Eds.), The handbook of research synthesis (pp. 97-109). New York: Russell Sage Foundation.

Wortman, P.M., Reichardt, C. S., & St. Pierre, R. G. (1978). The first year of the education voucher demonstration. Evaluation Quarterly, 2, 193-214.

Wortman, P.M., Smyth, J. M., Langenbrunner, J. C., & Yeaton, W. H. (1998). Consensus among experts and research synthesis. International journal of Technology Assessment in Health Care, 14, 109-122.

Wright, S. (1921). Correlation and causation. Journal of Agricultural Research, 20, 557-585.

Wright, S. (1934). The method of path coefficients. Annals of Mathematical Statistics, 5, 161-215.

Yaremko, R. M., Harari, H., Harrison, R. C., & Lynn, E. (1986). Handbook of research and quantitative methods in psychology for students and professionals. Hillsdale, NJ: Erlbaum.

Yeaton, W. H., & Sechrest, L. (1981). Critical dimensions in the choice and maintenance of successful treatments: Strength, integrity, and effectiveness. Journal of Consulting and Clinical Psychology, 49, 156-167.

Yeaton, W. H., & Sechrest, L. (1986). Use and misuse of no-difference findings in eliminating threats to validity. Evaluation Review, 10, 836-852.

Yeaton, W. H., & Wortman, P.M. (1993). On the reliability of meta-analytic reviews. Evaluation Review, 17, 292-309.

Yeaton, W. H., Wortman, P.M., & Langberg, N. (1983). Differential attrition: Estimating the effect of crossovers on the evaluation of a medical technology. Evaluation Review, 7, 831-840.

Yinger, J. (1995). Closed doors, opportunities lost. New York: Russell Sage Foundation.

Yu, J., & Cooper, H. (1983). A quantitative review of research design effects on response rates to questionnaires. Journal of Marketing Research, 20, 36-44.

Zabin, L. S., Hirsch, M. B., & Emerson, M. R. (1989). When urban adolescents choose abortion: Effects on education, psychological status, and subsequent pregnancy. Family Planning Perspectives, 21, 248-255.

Zadeh, L.A. (1987). Fuzzy sets and applications. New York: Wiley.

Zajonc, R. B., & Markus, H. (1975). Birrh order and intellectual development. Psychological Review, 82, 74-88.

Zeisel, H. (1973). Reflections on experimental technique in the law. Journal of Legal Studies, 2, 107-124.

Zelen, M. (1974). The randomization and stratification of patients to clinical trials. Journal of Chronic Diseases, 27, 365-375.

Zelen, M. (1979). A new design for randomized clinical trials? New England Journal of Medicine, 300, 1242-1245.

Zelen, M. (1990). Randomized consent designs for clinical trials: An update. Statistics in Medicine, 9, 645-656.

Zhu, S. (1999). A method to obtain a randomized control group where it seems impossible. Evaluation Review, 23, 363-377.

Zigler, E., & Weikart, D.P. (1993). Reply to Spitz's comments. American Psychologist, 48, 915-916.

Zigulich, J. A. (1977). A comparison of elementary school environments: Magnet schools versus traditional schools. Unpublished doctoral dissertation, Northwestern University, Evanston, IL.

Zucker, D. M., Lakatos, E., Webber, L. S., Murray, D. M., McKinlay, S. M., Feldman, H. A., Kelder, S. H., & Nader, P.R. (1995). Statistical design of rhe child and adolescent trial for cardiovascular health (CATCH): Implications of cluster randomization. Controlled Clinical Trials, 16, 96-118.