Diseños experimentales y cuasiexperimentales en la investigación social

Diseños experimentales y cuasiexperimentales en la investigación social

Donald T. Campbell
Julian C. Stanley

Amorrortu editores
Buenos Aires

Director de la biblioteca de sociología, Luis A. Rigal
Experimental and Quasi-Experimental Designs for Research, Donald T. Campbell y Julian C. Stanley

Primera edición en inglés, 1966; sexta reimpresión, 1970
Primera edición en castellano, 1973; primera reimpresión, 1974; segunda reimpresión, 1978; tercera reimpresión, 1982; cuarta reimpresión, 1988; quinta reimpresión, 1991; sexta reimpresión, 1993; séptima reimpresión, 1995
Traducción, Mauricio Kitaigorodzki
Revisión, José C. Orries e Ibars

Unica edición en castellano autorizada por *Rand McNally & Company*, Chicago, y debidamente protegida en todos los países. Queda hecho el depósito que previene la ley nº 11.723.

ISBN 950-518-042-X

Impreso en los Talleres Gráficos Color Efe, Paso 192, Avellaneda, provincia de Buenos Aires, en junio de 1995.

Tirada de esta edición: 2.000 ejemplares.

Nota preliminar

Este trabajo apareció originalmente en el libro compilado por N. L. Gage, *Handbook of research on teaching* (Chicago: Rand McNally Co., 1963), con un título algo distinto: «Diseños experimentales y cuasiexperimentales en la investigación educacional». Por esa razón, las primeras páginas y gran parte de los ejemplos ofrecidos versan sobre la investigación en el campo educativo. No obstante, si se examina la lista de referencias bibliográficas al final de la obra, se observará que el estudio que aquí presentamos extrae sus datos de todas las ciencias sociales, siendo por tanto de aplicación general sus recomendaciones metodológicas.

Donald T. Campbell
Julian C. Stanley

1. Introducción

Examinaremos en esta obra [1] la validez de dieciséis diseños experimentales respecto de doce amenazas corrientes a la inferencia válida. Por «experimento» entendemos aquella parte de la investigación en la cual se manipulan ciertas variables y se observan sus efectos sobre otras. Conviene aclarar que el propósito particular de este libro *no es* estudiar el diseño experimental dentro de la tradición de Fisher [1925, 1935], donde el experimentador, con pleno dominio de la situación, programa tratamientos y mediciones a fin de lograr la mejor eficiencia estadística, único objetivo al que obedece la mayor o menor complejidad del diseño. Los diseños aquí analizados son tanto más complejos cuanto mayor es la inflexibilidad del ambiente; es decir, en la medida en que el experimentador carece de control absoluto sobre la situación. Aunque hay no pocos puntos de contacto entre nuestro tratamiento y el de la corriente de Fisher, juzgamos apropiado dejar la exposición de esta última para obras de mayor envergadura, como las de Brownlee [1960], Cox [1958], Edwards [1960], Ferguson [1959], Johnson [1949], Johnson y Jackson [1959], Lindquist [1953], McNemar [1962] y Winer [1962]. (También puede consultarse Stanley, 1957*b*.)

1 La preparación de esta obra, en la que colaboraron Keith N. Clayton y Paul C. Rosenblatt, contó con el auspicio del Proyecto Psicología-Educación de la Northwestern University, bajo el patrocinio de la Carnegie Corporation.

2. El problema y sus antecedentes

McCall como modelo

En 1923, W. A. McCall publicó un libro titulado *How to experiment in education* (Cómo experimentar en educación). Nuestro propósito es exponer aquí en forma actualizada los intereses indicados y las consideraciones apuntadas en dicha obra; comenzaremos, pues, formulando una evaluación acerca de ella. Decía McCall en su introducción: «Hay excelentes libros y tratados que exponen el manejo estadístico de datos experimentales, pero muy pocos acerca de cómo obtener datos adecuados y correctos a los cuales poder aplicar el procedimiento estadístico». Este enunciado continúa siendo hoy tan cierto que bien puede servirnos de *leitmotiv*. Aunque la influencia de la corriente fisheriana remedió la situación en algunos aspectos fundamentales, su efecto más conspicuo parece haber sido el de afinar y perfeccionar el análisis estadístico, más que el de ayudar a conseguir «datos adecuados y correctos».

Quizá por su orientación práctica y sentido común, y porque no pretende constituir un aporte capital, el libro de McCall representa un clásico insuficientemente valorado todavía. Cuando apareció, dos años antes de la primera edición de *Statistical methods for research workers* (Métodos estadísticos para investigadores), de Fisher [1925], no había nada comparable cualitativamente a él en el ámbito de la agricultura ni en el de la psicología. Se anticipó en varios puntos fundamentales a las metodologías ortodoxas de esas dos ciencias. Acaso la más importante de las contribuciones de Fisher haya sido la idea de formular la igualación preexperimental de grupos por aleatorización.

Esta idea, y el consecuente rechazo de la tentativa de llegar a tal igualación por equiparación (pese a su intuitiva atracción y potencialidad de error), no mereció fácilmente la aprobación de los investigadores del ámbito educacional. En 1923, McCall había comprendido cuáles eran los elementos cualitativos fundamentales del problema. Dio, como primer método para establecer grupos comparables, el de los «grupos igualados por azar». «Así como se puede lograr la representatividad por el método aleatorio (...) también se puede conseguir la equivalencia por el mismo medio, siempre que el número de sujetos que hayan de utilizarse sea lo suficientemente grande» (pág. 41). También en otro punto se anticipó a Fisher: la introducción del diseño del cuadrado latino con el rótulo de «experimento rotatorio», que por otra parte habían utilizado ya Thorndike, McCall y Chapman [1916], tanto en formas 5×5 como 2×2, unos 10 años antes de que Fisher [1926] lo incorporase de modo sistemático a su esquema de diseño experimental con aleatorización.[2]

La forma en que McCall utiliza el «experimento rotatorio» ilustra muy bien el énfasis tanto de su obra como de la presente. El «experimento rotatorio» se introduce, no por razones de eficiencia, sino más bien para lograr algún control cuando no es posible la asignación aleatoria a grupos equivalentes. Con una intención similar examinaremos aquí las imperfecciones de muchos programas experimentales, abogando no obstante por su aplicación en aquellas configuraciones en que no haya modo de recurrir a mejores diseños experimentales. En este sentido, la mayor parte de los diseños analizados, incluso el «experimento rotatorio» no aleatorizado, se denominan diseños *cuasi*experimentales.

La desilusión provocada por los experimentos llevados a cabo en el campo de la educación

En esta obra nos declaramos partidarios del método experimental como único medio de zanjar las disputas relativas a la práctica educacional, única forma de verificar adelantos en el campo pedagógico y único método para acumular un saber al cual puedan introducírsele mejoras sin correr el peligro de que se descarten caprichosamente los conocimientos ya adquiridos a cambio de novedades de inferior calidad. Sin embargo, con nuestra enérgica defensa de la experimentación no pretendemos significar que este énfasis sea nuevo. Como lo manifiesta la existencia misma del libro de McCall, en tiempos de Thorndike una ola de entusiasmo experimental recorría el ámbito de la educación, alcanzando quizá su punto culminante

2 Kendall y Buckland [1957] afirman que el cuadrado latino fue inventado por el matemático Euler en 1782. Thorndike, Chapman y McCall no utilizan esta expresión.

en la década del veinte. Aquel entusiasmo se convirtió después en apatía y rechazo, así como en la adopción de nuevas doctrinas psicológicas no susceptibles de verificación experimental. Good y Scates [1954, págs. 716-21] han documentado un pesimismo general, que se retrotrae quizás a 1935, y citan incluso a Monroe [1938], aquel decidido defensor de la experimentación controlada nos han desilusionado». Cabe destacar, además, que el tránsito de la experimentación a la redacción de ensayos, acompañado a menudo por una conversión del conductismo tipo Thorndike a la psicología de la guestalt o al psicoanálisis, se produjo con frecuencia en personas que contaban con una buena formación en la tradición experimental.

Para evitar que se repita este desencanto, debemos conocer los orígenes de la reacción anterior, procurando sortear las falsas expectativas que condujeron a ella. Merecen destacarse varios aspectos. Ante todo, se pretendió asignar a los resultados de la experimentación un cierto ritmo y grado exagerado de progreso, al par que se menospreciaba injustificadamente el conocimiento no experimental. Los primeros defensores supusieron que el progreso en la tecnología pedagógica había sido lento *solo porque* no se había aplicado a ella el sistema científico: creían que la práctica tradicional era ineficaz solo porque no había sido fruto de la experimentación. Cuando se demostró que los experimentos eran a menudo tediosos, equívocos, de reiterabilidad insegura y ratificadores, por lo común, de conocimientos precientíficos, los fundamentos excesivamente optimistas con que se había querido justificar la experimentación quedaron minados por la base, y al primitivo entusiasmo sucedió el desilusionado abandono.

Aquella sensación era compartida tanto por los observadores como por los propios involucrados. Entre los experimentadores se advertía una innegable aversión hacia la experimentación. Para el investigador normal, muy motivado, el hecho de que una de las hipótesis que sustenta no sea confirmada resulta por demás doloroso. Como animal biológico y psicológico, está sujeto a leyes de aprendizaje que lo conducen inevitablemente a asociar este dolor con los estímulos y acontecimientos inmediatos. No es extraño, pues, que tales estímulos estén constituidos por el mismo proceso experimental de modo más vívido y directo que la «verdadera» fuente de la frustración, a saber: la inadecuada teoría. Una situación tal puede inducir, inconscientemente quizás, a evitar o rechazar el proceso experimental. Si, como parece probable, la ecología

de nuestra ciencia está constituida de tal manera que hay en ella muchas más respuestas erróneas que correctas, cabe prever el fracaso de la mayor parte de los experimentos. Hay que inmunizar, pues, de algún modo a los jóvenes investigadores contra ese resultado y, en general, justificar ante ellos la experimentación sobre fundamentos más realistas: no como una panacea, pero sí como el único camino hacia el progreso acumulativo. Tenemos que inculcar en nuestros discípulos la expectativa del tedio y la decepción, y el deber de la tenaz persistencia, actitudes ambas que con tanto éxito se ha logrado implantar ya en las ciencias biológicas y físico-naturales. Hay que ampliar el voto de pobreza de nuestros alumnos, de modo que no solo se avengan a trabajar con insuficientes recursos financieros sino a admitir la insuficiencia de sus resultados experimentales.

Más concretamente: debemos ensanchar nuestra perspectiva temporal, y reconocer que la experimentación continua y múltiple es más propia de la actividad científica que los experimentos únicos y definitivos. Las pruebas que realizamos hoy, si llegan a tener éxito, exigirán repetición y validaciones cruzadas en otros momentos y en otras condiciones antes de convertirse en adquisición estable para el acervo científico y ser susceptibles de segura interpretación teórica. Además, aun cuando reconocemos que la experimentación es el lenguaje fundamental de la demostración y el único tribunal decisivo para resolver los desacuerdos entre posibles teorías rivales, no es previsible que los «experimentos cruciales» que contrapongan a las teorías opuestas vayan a producir resultados claramente definitorios. Cuando se descubra, por ejemplo, que observadores competentes sustentan puntos de vista muy dispares entre sí, será razonable suponer *a priori* que ambos habrán encontrado algo válido sobre la situación estudiada, y que ambos representarán una parte de la verdad completa. Cuanto mayor sea la controversia, más probable será que así ocurra. Podemos, pues, esperar en tales casos un resultado experimental de carácter mixto, o con sutiles variaciones en el saldo de verdad entre una prueba y otra. La posición más sensata —lograda en gran parte por la psicología experimental (por ejemplo, Underwood, 1957*b*)— evita los experimentos cruciales, reemplazándolos por relaciones e interacciones dimensionales a lo largo de muchas gradaciones diversas de las variables.

Tampoco hay que olvidar los muy perfeccionados procedimientos estadísticos que en época reciente se han ido introduciendo poco a poco en la psicología y la educación. Durante su período

de mayor actividad, la experimentación educacional avanzó lentamente, empleando medios y procedimientos burdos. McCall [1923] y sus contemporáneos realizaron investigaciones en las cuales se estudiaba una sola variable por vez. Para la enorme complejidad que caracteriza las situaciones de aprendizaje humano, aquello resultaba demasiado lento. Hoy se sabe la gran importancia que pueden asumir diversas contingencias, dependientes de la «acción» conjunta de dos o más variables experimentales. Stanley [1957*a*, 1960, 1961*b*, 1961*c*, 1962], Stanley y Wiley [1962] y otros han destacado la imperiosa necesidad de evaluar tales interacciones.

Los experimentos pueden incluir algunas variables en cualquiera de dos sentidos o en ambos a la vez. Por ejemplo, incorporando al diseño más de una variable «independiente» (sexo, grado escolar, método con que se enseña aritmética, estilo y tamaño de los tipos de imprenta, etc.), y/o empleando más de una variable «dependiente» (número de errores, velocidad, diversas pruebas, etc.). Los procedimientos de Fisher son multivariados en el primer sentido y univariados en el segundo. Estadísticos matemáticos como Roy y Gnanadesikan [1959] tratan de encontrar diseños y análisis que unifiquen ambas formas de diseños multivariados. Tal vez permaneciendo alertas a la evolución de tales diseños puedan los investigadores en el campo de la educación reducir la brecha, por lo común demasiado amplia, entre la exposición en la literatura especializada de un procedimiento estadístico y su aplicación práctica a investigaciones de envergadura.

No cabe duda de que una capacitación más a fondo de los investigadores educacionales en técnicas *modernas* de estadística experimental permitiría elevar la calidad de la experimentación pedagógica.

Concepción evolutiva sobre la ciencia y la acumulación de conocimientos

Como fundamento de lo expuesto en los párrafos precedentes y lo que se expondrá en los que siguen señalamos una concepción evolutiva del conocimiento [Campbell, 1959], según la cual la aplicación práctica y el conocimiento científico son el resultado de la acumulación de ciertas tentativas seleccionadas y remanentes del caudal de observaciones recogidas por la experiencia. Esta concepción inspira gran respeto por la tradición en la práctica pedagógica. Si en el trascurso de los siglos se han ensayado muchos enfoques distintos, si de ellos algunos han obtenido mejores resultados que otros y los que mejor funcionaban es de suponer que habrán sido los aplicados con mayor persistencia por sus creadores, imitados por otros y trasmitidos a las generaciones siguientes, las costumbres resultantes de todo ello pueden representar un valioso y probado subconjunto de todas las prácticas posibles.

Pero el punto de corte selectivo de esta evolución se torna muy impreciso cuando se lo traslada a la realidad. Las condiciones de observación, tanto físicas como psicológicas, distan mucho de ser óptimas. Lo que sobrevive o se retiene queda en gran parte determinado por el azar. Es aquí donde la experimentación demuestra la importancia del proceso de prueba, exploración y selección. No se contempla, pues, la experimentación en sí misma como fuente de ideas necesariamente contradictorias con relación al saber tradicional, sino más bien como mecanismo de refinación superpuesto a las acumulaciones probablemente valiosas de la práctica sensata. Propugnar, pues, una ciencia experimental de la educación no implica repudiar el saber tradicional.

Algunos lectores abrigarán tal vez la sospecha de que la analogía con el esquema evolutivo darwiniano se complique con factores de carácter específicamente humano. Cuando Juan Pérez, director de escuela, tiene que decidir entre adoptar un libro de texto modificado o continuar con la versión anterior, es probable que haga su elección fundándose en datos insuficientes. Aparte de la eficiencia misma para la enseñanza y el aprendizaje, son muchas las consideraciones que habrá de tomar en cuenta. El director hará lo correcto en una de estas dos formas posibles: reteniendo el libro antiguo cuando sea tan bueno o mejor que el revisado, o adoptando este último cuando sea superior al primero. Pero puede equivocarse también de dos maneras: reteniendo el libro antiguo cuando el nuevo es mejor, o adoptando este cuando no es superior al primero. En cada una de las dos elecciones erróneas es de suponer que se producirán inconvenientes diversos: 1) mayor costo financiero y de gasto de energías; 2) costo para el director, en forma de quejas de los maestros, padres y miembros del consejo escolar; 3) costo para los maestros, los alumnos y la sociedad a causa de una peor instrucción. Estos costos, evaluados en términos de dinero, energía, confusión, menor aprendizaje y mayor riesgo personal, deben sopesarse frente a la probabilidad de que se produzca cada una de dichas alternativas, así como la de que se detecte el error mismo. Si el director toma

su decisión sin suficientes elementos de juicio, fruto de una investigación a fondo, sobre el costo 3 (peor instrucción), es posible que exagere los costos 1 y 2. Los naipes vienen barajados en favor de un criterio conservador: mantener el libro antiguo durante un año más. Cabe, sin embargo, tratar de preparar un experimento con ambos libros a la vez, de acuerdo con un esquema de teoría de la decisión [Chernoff y Moses, 1959], y adoptar una resolución que tome explícitamente en cuenta los diversos costos y probabilidades. Cómo conseguir que las cuidadosas deliberaciones de un excelente administrador educativo se aproximen a este modelo de teoría de la decisión es un grave problema, cuyo estudio bien vale la pena encarar.

Factores que atentan contra la validez tanto interna como externa

En los próximos capítulos de esta obra se describen doce factores que amenazan la validez de varios diseños experimentales.[3] Cada uno de dichos factores se explicará con todo detalle al exponer los diseños a propósito de los cuales constituye un problema particular; diez de los dieciséis diseños se presentarán antes de completarse la lista. A fin de lograr una perspectiva más clara sería conveniente, sin embargo, que demos una lista de dichos factores, así como una guía general acerca de los cuadros 1, 2 y 3, que resumen parcialmente el análisis. Es fundamental a este respecto distinguir bien entre *validez interna* y *validez externa*. Llamamos *validez interna* a la mínima imprescindible, sin la cual es imposible interpretar el modelo: ¿Introducían, en realidad, una diferencia los tratamientos empíricos en este caso experimental concreto? Por su parte la *validez externa* plantea el interrogante de la *posibilidad de generalización*: ¿A qué poblaciones, situaciones, variables de tratamiento y variables de medición puede generalizarse este efecto? Ambos criterios son sin duda importantes, aunque con frecuencia se contrapongan, en el sentido de que ciertos aspectos que favorecen a uno de ellos perjudican al otro. Si bien la *validez interna* es el *sine qua non*, y a la cuestión de la *validez externa*, como a la de la inferencia inductiva, nunca se puede responder plenamente, es obvio que nuestro ideal lo constituye la selección de diseños ricos en una y otra validez. Así ocurre, particularmente, respecto de la investigación sobre métodos de enseñanza, donde el desiderátum será la generalización a situaciones prácticas de carácter conocido. Tanto las distinciones como las relaciones entre estos dos tipos de consideraciones de validez irán haciéndose más explícitas a medida que se las ilustre durante la exposición de diseños específicos.

Con relación a la *validez interna*, presentaremos ocho clases distintas de variables externas que, de no controlárselas en el diseño experimental, podrían generar efectos que se confundirían con el del estímulo experimental. Constituyen los efectos de:

1. *Historia*, los acontecimientos específicos ocurridos entre la primera y la segunda medición, además de la variable experimental.
2. *Maduración*, procesos internos de los participantes, que operan como resultado del mero paso del tiempo (no son peculiares de los acontecimientos en cuestión), y que incluyen el aumento de la edad, el hambre, el cansancio y similares.
3. *Administración de tests*, el influjo que la administración de un test ejerce sobre los resultados de otro posterior.
4. *Instrumentación*, los cambios en los instrumentos de medición o en los observadores o calificadores participantes que pueden producir variaciones en las mediciones que se obtengan.
5. *Regresión estadística*, opera allí donde se han seleccionado los grupos sobre la base de sus puntajes extremos.
6. Sesgos resultantes en una *selección* diferencial de participantes para los grupos de comparación.
7. *Mortalidad experimental*, o diferencia en la pérdida de participantes de los grupos de comparación.
8. *Interacción entre la selección y la maduración*, etc., en algunos de los diseños cuasiexperimentales de grupo múltiple, como el 10, se confunde con el efecto de la variable experimental (es decir, que podría tomarse por él).

Los factores que amenazan la *validez externa* o *representatividad*, y que vamos a analizar aquí, son:

9. El *efecto reactivo* o *de interacción* de las *pruebas*, cuando un pretest podría aumentar o disminuir la sensibilidad o la calidad de la reacción del participante a la variable experimen-

3 Gran parte de esta exposición se funda en Campbell [1957]. En general, no se harán referencias particulares a esta fuente.

tal, haciendo que los resultados obtenidos para una población con pretest no fueran representativos de los efectos de la variable experimental para el conjunto sin pretest del cual se seleccionaron los participantes experimentales.
10. Los efectos de *interacción* de los sesgos de *selección* y la *variable experimental*.
11. *Efectos reactivos de los dispositivos experimentales*, que impedirían hacer extensivo el efecto de la variable experimental a las personas expuestas a ella en una situación no experimental.
12. *Interferencias de los tratamientos múltiples*, que pueden producirse cuando se apliquen tratamientos múltiples a los mismos participantes, pues suelen persistir los efectos de tratamientos anteriores. Este es un problema particular de los diseños de un solo grupo de tipo 8 o 9.

En la presentación de los diseños experimentales se adoptarán un código y unos símbolos gráficos uniformes, a fin de compendiar la mayoría, si no la totalidad, de sus características distintivas. Una *X* representará la exposición del grupo a una variable o acontecimiento experimental, cuyos efectos se han de medir; *O* hará referencia a algún proceso particular de observación o medición; las *X* y *O* en una fila dada se aplican a las mismas personas específicas. La dimensión representada de izquierda a derecha indica el orden temporal, en tanto que las *X* y *O* dispuestas en forma vertical señalan la presencia de simultaneidad. Para hacer ciertas distinciones importantes, como entre los diseños 2 y 6 o entre el 4 y el 10, hay que utilizar un símbolo *R*, que indica asignación aleatoria a diferentes grupos de tratamiento. Esa aleatorización se concibe como un proceso que se produce en un momento dado, y sirve para lograr, dentro de límites estadísticos conocidos, la igualdad de los grupos antes del tratamiento. Agregaremos a ella otra convención gráfica: las filas paralelas no separadas por línea de puntos significan grupos de comparación no igualados por dicho procedimiento. No se ha empleado ningún símbolo para la equiparación como proceso para conseguir la igualación previa al tratamiento de grupos de comparación, porque el valor de dicho proceso se ha exagerado mucho y suele más bien conducir a inferencias erróneas que contribuir a extraer conclusiones válidas. (Véanse más adelante el análisis del diseño 10 y la sección final sobre diseños correlacionales). En el diseño 9 se ha utilizado explícitamente un símbolo *M* para identificar materiales.

3. Tres diseños preexperimentales

1. Estudio de caso con una sola medición

Gran parte de las investigaciones actuales sobre educación se ajustan a un diseño en el cual se estudia un solo grupo cada vez, después de someterlo a la acción de algún agente o tratamiento que se presuma capaz de provocar un cambio. Estos estudios podrían diagramarse de la siguiente forma:

X O

Como ya se ha destacado [p. ej., Boring, 1954; Stouffer, 1949], tales estudios adolecen de tan absoluta falta de control que su valor científico es casi nulo. Presentamos este diseño como punto mínimo de referencia. No obstante, a causa de la continua inversión en esta clase de estudios y de la extracción de inferencias causales de ellos, será imprescindible formular alguno que otro comentario. El proceso de comparación, de registro de diferencias o de contrastes es fundamental para la comprobación científica (y para todos los procesos de diagnóstico del conocimiento, incluso aquellos vinculados con la retina). Resulta ilusoria cualquier apariencia de conocimiento absoluto o intrínseco sobre objetos singulares aislados. La obtención de datos científicos implica, por lo menos, una comparación, cuya utilidad depende de que las partes integrantes se estructuren con el mismo cuidado e idéntica precisión.
En los estudios de casos del diseño 1, se compara implícitamente un caso único, cuidadosamente estudiado, con otros acontecimientos observados de manera casual y recordados. Las inferencias se fundan en expectaciones generales de cuáles hubieran sido los datos de no haberse producido *X*, etc. Tales estudios suelen requerir una tediosa recopilación de detalles concretos, cuidadosa observación, administración de tests y similares, y en tales casos se corre el riesgo de hacer *precisiones injustificadas*. ¡Cuánto más provechoso sería el estudio si ese caudal de observaciones se redujese a la mitad, aplicándose

el esfuerzo ahorrado al estudio igualmente cuidadoso de un apropiado caso de comparación! Parece hasta casi falto de ética el aceptar hoy, como tesis de doctorado en el ámbito educacional, estudios de casos de esa índole (es decir, que implican un solo grupo observado una sola vez). En ellos, los tests «estandarizados» solo ofrecen una ayuda muy limitada, puesto que las fuentes antagónicas de diferencias (distintas de X) son tan abundantes que tornan casi inútil el grupo «estándar» de referencia como «grupo de control». Por los mismos motivos, las muchas fuentes no controladas de diferencias entre el estudio actual de un caso concreto y otros que, planteándose en el futuro, pudieran compararse con aquel son tantas, que hacen también inútil su justificación como punto de referencia para estudios posteriores. En general, sería mejor distribuir el esfuerzo descriptivo entre los dos miembros de una comparación interesante.

Si se lo toma en conjunto con las comparaciones implícitas de «conocimiento común», el diseño 1 presenta la mayor parte de los inconvenientes de cada uno de los diseños posteriores. Por eso dejaremos el estudio de esos inconvenientes para cuando encaremos situaciones más específicas.

2. Diseño pretest-postest de un solo grupo

Si bien este diseño continúa siendo de gran aplicación en la investigación educacional, y se lo considera tan superior al diseño 1 que se lo utiliza allí donde no cabe hacer nada mejor (véase más adelante el análisis de los diseños cuasiexperimentales), lo presentamos aquí como un «mal ejemplo» para ilustrar algunas de las variables externas entremezcladas que pueden atentar contra la validez *interna*. Esas variables ofrecen hipótesis aceptables que explican una diferencia $O_1 - O_2$, opuesta a la hipótesis de que X causó la diferencia:

$$O_1 \quad X \quad O_2$$

La primera de estas hipótesis rivales no controladas es la *historia*. Entre O_1 y O_2 pueden haber ocurrido muchos otros acontecimientos capaces de determinar cambios, además de la X sugerida por el experimentador. Si el pretest (O_1) y el postest (O_2) se administraron en días distintos, los acontecimientos intermedios pueden haber causado la diferencia. Para

convertirse en una hipótesis rival *aceptable*, tal acontecimiento debería haber afectado a la mayor parte de los estudiantes que integran el grupo examinado (p. ej., en algún otro período lectivo o por medio de una noticia periodística muy difundida). En el estudio escolar realizado por Collier en 1940, sobre el cual informó en 1944, se produjo la caída de Francia mientras los estudiantes leían abundante material de propaganda nazi; los cambios de actitud comprobados parecieron ser consecuencia, más probablemente, de ese suceso que de la propaganda en sí.[1] La *historia* se convierte en una explicación rival más aceptable del cambio cuanto más extenso es el lapso entre O_1 y O_2, y podría considerarse un detalle trivial en un experimento realizado dentro del breve lapso de una o dos horas, si bien aun en tal caso deben investigarse fuentes externas como las risas, las distracciones, etc. La variable *historia* se relaciona con la característica de *aislamiento experimental*, que en muchos laboratorios de física suele conseguirse con tanta aproximación que el diseño 2 resulta aceptable a propósito de la mayor parte de sus investigaciones. Pero en el estudio de métodos de enseñanza casi nunca se puede suponer un aislamiento experimental tan completo. Por eso en el cuadro 1 el diseño 2 se ha marcado con un signo negativo bajo el título *Historia*, en el que incluiremos un grupo de posibles efectos estacionales o de programación de acontecimientos institucionales, aunque también estos podrían situarse al pie del título *Maduración*. Así, el optimismo podría variar con las estaciones y la ansiedad producida por el programa de exámenes semestrales [p. ej., Crook, 1937; Windle, 1954]. Tales efectos acaso produjesen una variación $O_1 - O_2$ confundible con el efecto de X.

Una segunda variable o categoría de variables rivales recibe el nombre de *maduración*. Tal como lo entendemos aquí, este término abarca todos aquellos procesos biológicos o psicológicos que varían de manera sistemática con el correr del tiempo e independientemente de determinados acontecimientos externos. Así, es probable que entre O_1 y O_2 los estudiantes hayan aumentado de edad, apetito, fatiga, aburrimiento, etc., y acaso la diferencia obtenida refleje ese cambio y no el de X. En educación correctiva, que se aplica a personas excepcionalmente disminuidas, un proceso de «remisión espontánea», análogo al que se produce en la curación de heridas,

1 En realidad, Collier utilizó un diseño más adecuado que este, que en el presente sistema se denomina diseño 10.

Cuadro 1. *Fuentes de invalidación para los diseños 1 a 6.*

	Fuentes de invalidación										
	Interna								Externa		
	Historia	Maduración	Administración de tests	Instrumentación	Regresión	Selección	Mortalidad	Interacción de selección y maduración, etc.	Interacción de administración de tests y *X*	Interacción de selección y *X*	Dispositivos reactivos	Interferencia de *X* múltiples
Diseños preexperimentales												
1. Estudio de caso con una sola medición *X O*	—	—				—	—			—		
2. Diseño pretest-postest de un solo grupo *O X O*	—	—	—	—	?	+	+	—	—	—	?	
3. Comparación con un grupo estático *X O* *O*	+	?	+	+	+	—	—	—		—		
Diseños experimentales propiamente dichos												
4. Diseño de grupo de control pretest-postest *R O X O* *R O O*	+	+	+	+	+	+	+	+	—	?	?	
5. Diseño de cuatro grupos de Solomon *R O X O* *R O O* *R X O* *R O*	+	+	+	+	+	+	+	+	+	?	?	
6. Diseño de grupo de control con postest únicamente *R X O* *R O*	+	+	+	+	+	+	+	+	+	?	?	

Nota: En los cuadros, el signo negativo indica que hay imperfección definida; el positivo, que el factor está controlado; el interrogativo, la presencia de una posible causa de preocupación, y por último, el espacio en blanco significa que el factor no es pertinente.
Estos cuadros resumidos los presentamos con suma renuencia, ya que pueden resultar «demasiado útiles», si se llega a confiar en ellos y no en la exposición más completa y calificada que se incluye en el texto. Ningún indicador de + o — debe respetarse, a menos que el lector comprenda por qué se lo ha colocado. En particular, va contra el espíritu de este trabajo la creación de una confianza o suspicacia infundadas con respecto a determinados diseños.

puede confundirse con el efecto específico de una *X* correctiva. (Ni que decir tiene que tal remisión no se considera «espontánea» en ningún sentido causal, sino que representa más bien los efectos acumulativos de los procesos de aprendizaje y presiones ambientales de la experiencia global diaria, que se producirían aunque no se hubiese introducido ninguna *X*.)
Una tercera explicación rival entremezclada es el efecto de la *realización de pruebas*, el efecto del pretest mismo. En pruebas de rendimiento e inteligencia, los estudiantes a quienes se somete a ellas por segunda vez, o a una de sus variantes, etc., suelen desempeñarse mejor que los que las encaran por vez primera [p. ej., Anastasi, 1958, págs. 190-91; Cane y Heim, 1950]. Esos efectos, que alcanzan de 3 a 5 puntos de CI en promedio para sujetos sin experiencia previa, se producen aun sin haberles hecho comentario alguno acerca de sus puntajes o errores en el test anterior. En las pruebas de personalidad se advierte un resultado similar: en las segundas se observa, en general, un mejor ajuste, aunque en ocasiones se halla también un efecto altamente significativo en sentido contrario [Windle, 1954]. En cuanto a las actitudes hacia grupos minoritarios, una segunda prueba suele indicar un mayor prejuicio, aunque los datos disponibles son todavía escasos [Rankin y Campbell, 1955]. Es obvio que el anonimato, una mayor conciencia de qué respuesta es la socialmente aprobada, etc., influirían en general sobre la índole del resultado. Para tests de prejuicio en condiciones de anonimato, el nivel de adaptación creado por las expresiones hostiles presentadas puede modificar las apreciaciones del estudiante en lo referente a la tolerancia que existe para actitudes de mayor hostilidad. En un inventario de adaptación o de personalidad que lleva la firma del sujeto, la primera administración del test forma parte de una situación de solución de problemas en que el estudiante trata de descubrir el propósito oculto de la prueba. Si ya ha pasado por aquella experiencia (o si habló con sus amigos sobre las respuestas que ellos dieron a algunos de los puntos más destacados), sabe mejor cómo comportarse la segunda vez.
Con el problema de los efectos del test se relaciona la distinción entre las posibles mediciones de su *reactividad*, lo cual constituirá un importante tema en todo este libro, así como una exhortación general a que se hagan mediciones *no reactivas* siempre que sea posible. Desde hace mucho tiempo ha sido una verdad manifiesta en las ciencias sociales que el proceso mismo de medición puede hacer cambiar aquello que se

mide. La ganancia test-retest sería una importante consecuencia de ese cambio. (Otra, la interacción entre la realización de la prueba y X, la estudiaremos más adelante, junto con el diseño 4. Además, es importante evitar esas reacciones al pretest, aun cuando surtan efectos diferentes para sujetos distintos.) Es de esperar el efecto reactivo siempre que el proceso de prueba sea en sí un estímulo al cambio, y no un mero registro de comportamiento. Así, en un experimento sobre terapia para el control del peso, el pesaje inicial puede ser de suyo un estímulo para el adelgazamiento, aun sin tratamiento curativo alguno. De manera similar, la ubicación de observadores en el aula para estudiar la capacidad preentrenamiento del docente en el ámbito de las relaciones humanas puede modificar de por sí su forma de comportarse. La colocación de un micrófono sobre el escritorio o pupitre suele variar la pauta de interacción del grupo, etc. En general, cuanto más nuevo y motivante sea el elemento utilizado para las pruebas, mayor será su influencia.

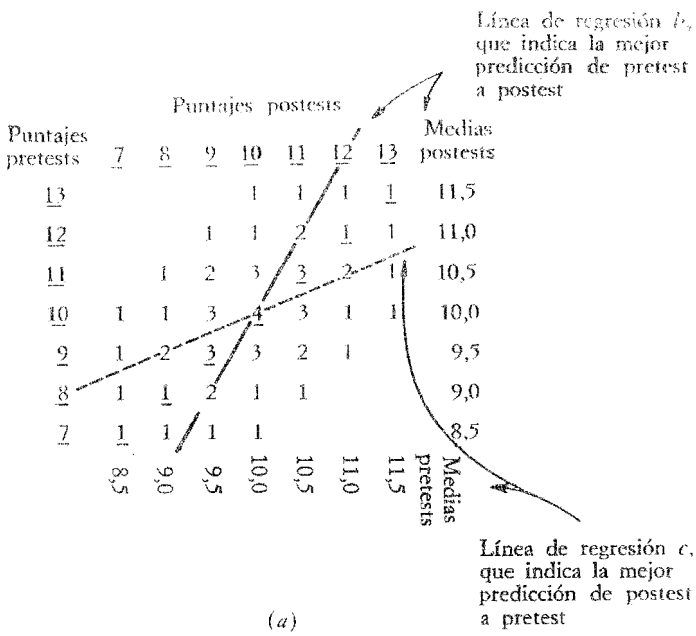
La *instrumentación* o «deterioro de los instrumentos» [cf. Campbell, 1957] es el término con que se designa una cuarta hipótesis rival no controlada. Esa expresión se refiere a las variaciones autónomas en el instrumento de medición que podrían ser la causa de una diferencia $O_1 - O_2$. Tales cambios serían análogos a la mayor o menor tensión observada en el dinamómetro, la condensación en una cámara de niebla, etc. Cuando se recurre a observadores humanos a fin de obtener O_1 y O_2, su propio aprendizaje, tensión, etc., determinarán diferencias de $O_1 - O_2$. Si se califican los ejercicios de redacción, ensayos o trabajos de investigación, los estándares aplicados variarán de O_1 a O_2 (la técnica de control sugiere que se mezclen los ejercicios de redacción O_1 y O_2 y se los haga calificar sin tener conocimiento de cuál ha llegado primero). Si se observa la participación en el aula, tal vez en la segunda sesión los observadores sean más hábiles, o más indiferentes. Si se entrevista a los padres, la familiaridad de quien realiza esa labor con el programa de entrevistas y con determinados padres puede producir ciertos desplazamientos. Un cambio en los observadores entre O_1 y O_2 también podría provocar alguna diferencia.

Una quinta variable entremezclada en algunos casos del diseño 2 es la *regresión estadística*. Por ejemplo, si en una prueba correctiva se seleccionan alumnos para un experimento especial porque han tenido puntajes particularmente bajos en el test de rendimiento escolar (que para ellos se convierte en

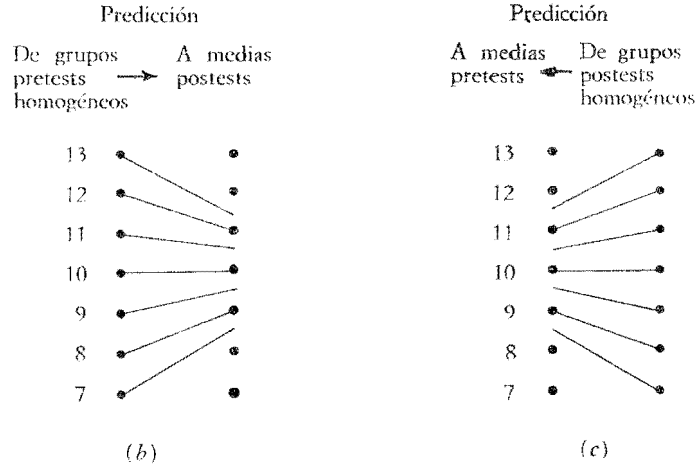
O_1), en una prueba posterior en que se adopte la misma forma de antes u otra similar a ella, casi con seguridad O_2 tendrá para ese grupo un promedio más elevado que O_1. Este resultado confiable no se deberá a ningún efecto genuino de X, a ningún efecto de la práctica de test y retest, etc. Es más bien un aspecto tautológico de la correlación imperfecta entre O_1 y O_2. Los errores de inferencia ocasionados por no haber tomado en cuenta el efecto de la regresión han planteado tantos problemas en la investigación educacional porque muy a menudo se desconoce su verdadera naturaleza —aun por estudiantes que han realizado cursos avanzados de estadística moderna—. Como en exposiciones posteriores (p. ej., el diseño 10 y el análisis *ex post facto*) la daremos por conocida, nos detendremos aquí a explicarla brevemente, aunque sea en forma muy elemental. La figura 1 presenta algunos datos imaginarios en los que el pretest y el postest de una población entera tienen una correlación de 0,50, sin variación en la media grupal o variabilidad. (Los datos se seleccionaron expresamente para que la colocación de las medias de fila y columna sean obvias a la simple observación visual. El valor de 0,50 también se elige por conveniencia de exposición.) En este caso hipotético no se ha producido ningún cambio real, pero, como es corriente, los puntajes falibles del test indican una correlación de retest considerablemente inferior a la unidad. Si, como se sugirió en el ejemplo dado antes, comenzamos por observar solo a los escolares calificados con puntajes muy bajos en el pretest —p. ej., 7 puntos—, y en el postest solo reparamos en el puntaje de esos alumnos, nos encontraremos con que los puntajes postest están dispersos, pero son en general mejores, y en promedio «regresionaron» hacia la media grupal con un coeficiente de regresión o correlación de 0,50, obteniendo una media de 8,5. No obstante, en vez de constituir una prueba de progreso, esto es una ratificación tautológica, si bien específica, de que hay una correlación imperfecta, y de cuál es su medida.

Cuando al trascurrir el tiempo se producen acontecimientos entre el pretest y el postest, nos sentimos tentados a establecer una relación causal entre dicho cambio y la acción específica del paso del tiempo. Pero obsérvese que cabe hacer aquí un análisis cronológico a la inversa, comenzando, por ejemplo, con aquellos cuyo puntaje postest es 7 y observando la dispersión de sus puntajes pretest, de los cuales se extraería la implicación inversa, a saber: que los puntajes van empeorando.

Figura 1. *Regresión en la predicción de puntajes postest del pretest, y viceversa.*



(*a*)



(*b*)

(*c*)

Las inferencias causales más erróneas son las que se extraen cuando la información se presenta en la forma indicada en la figura 1(*b*) [o la parte superior o inferior de 1(*b*)]. Así se da la impresión de que los alumnos más brillantes van perdiendo su ventaja, y viceversa, como si fuese por el efecto vulgarizador y homogeneizante del medio institucional. Aunque esta errónea interpretación implica que la variabilidad poblacional en el postest debería ser menor que en el pretest, ambas son en realidad iguales. Más todavía: si se procede al análisis con grupos puros de puntajes postest [como en la línea de regresión *c* y la figura 1(*c*)], quizá se llegue a la conclusión contraria. Como lo señaló McNemar [1940], el uso del análisis de control de tiempo invertido y el examen directo en busca de cambios en las variabilidades poblacionales son precauciones útiles contra dicho error de interpretación.

Cabe observar la regresión hacia la media en otra forma análoga. Cuanto más desviado sea el puntaje, mayor será el probable error de medición. Así, en cierto sentido, el típico alumno habituado a la obtención de puntajes elevados se habrá visto favorecido por una «suerte» extraordinaria (gran error positivo), al paso que la mala fortuna acompañó a quien obtuvo puntaje muy bajo (gran error negativo). La suerte es, sin embargo, caprichosa, por lo cual en un postest se espera que quienes poseen puntajes elevados declinen algo en el promedio, así como que los de puntajes bajos mejoren su posición relativa. (Se aplica la misma lógica si se comienza con los puntajes de postest y se procede hacia atrás, en dirección al pretest.)

La regresión hacia la media es un fenómeno general, que no se limita a la administración del pretest y del postest con la misma prueba o formas similares de ella. El director que observa que sus estudiantes de mayor CI suelen obtener puntajes inferiores a los máximos (aunque muy elevados) en las pruebas de rendimiento escolar, mientras los de menor CI no suelen ocupar el extremo inferior en esas pruebas (aunque sí puestos bastante bajos), sería culpable de falacia en la regresión si dijese que su escuela subestimula a los alumnos más brillantes y recarga de trabajo a los atrasados. Si seleccionase a los que obtuvieron el mayor y el menor puntaje en la prueba de rendimiento y analizara sus CI, la misma falta de lógica lo forzaría a llegar a la conclusión opuesta.

Si bien hemos hablado aquí de la regresión a propósito de los errores de medición, en general, ella depende más bien del grado de correlación: cuanto menor sea esta, mayor será la

regresión hacia la media. La falta de correlación perfecta puede deberse a «error» y/o a fuentes sistemáticas de variancia específica propia de una o de otra medición.

Los efectos de la regresión son, pues, acompañamientos inevitables de la correlación imperfecta de test-retest para grupos *seleccionados por su ubicación extrema.* No son, sin embargo, concomitantes necesarios de puntajes extremos dondequiera que ellos se produzcan. Si un grupo *seleccionado por razones independientes* resulta poseer una media extrema, hay una menor expectación *a priori* de que la media grupal regresione en una segunda prueba, pues se ha permitido a las fuerzas aleatorias o externas de variancia que influyan sobre los puntajes iniciales en ambas direcciones. Pero no ocurre así en un grupo seleccionado *a causa* de su extremidad en una variable falible, pues ella es artificial y dicho grupo regresionará hacia la media de la población de donde se lo seleccionó.

Efectos de regresión más indirectos pueden obedecer a la selección de sujetos con puntos extremos en mediciones diferentes del pretest. Consideremos un caso en el cual se eligen, para recibir adiestramiento experimental, estudiantes que «fracasan» en pruebas tomadas en el aula. Como pretest, se les administra el tipo A de un test estándar de rendimiento escolar, y como postest el tipo B de dicho test. Es probable que la prueba tomada en clase tenga una correlación más alta con la administración inmediata del tipo A que con la administración del tipo B unos tres meses después (si en cada sesión toda la clase ha sido objeto de la prueba). Cuanto más elevada sea la correlación, menor será la regresión hacia la media. Por consiguiente, los fracasos de la clase habrán determinado una regresión ascendente menor en el pretest que en el postest, dando una seudoganancia que podría haberse confundido con un conato afortunado de educación correctiva. [Para más detalles sobre ganancias y regresión, véase Lord, 1956; McNemar, 1958; Rulon, 1941; R. L. Thorndike, 1942.]

Con ello se concluye la lista de inconvenientes del diseño 2 que podemos analizar en este momento. En el cuadro 1 aparece otro signo negativo bajo el título «Validez interna», correspondiente a un factor que no analizaremos hasta exponer el diseño 10 (véase página 93) en la sección de diseños cuasiexperimentales, y dos signos negativos bajo «Validez externa», que no explicaremos hasta haber realizado el análisis del diseño 4 (véase página 32).

3. Comparación con un grupo estático

El tercer diseño preexperimental necesario para nuestra exposición de los factores de invalidación es la comparación con un grupo estático. Es un diseño en el cual un grupo que ha experimentado X se compara con otro que no lo ha hecho, a fin de establecer el efecto de X.

$$\begin{array}{ll} X & O_1 \\ \hline & O_2 \end{array}$$

Ejemplos de esta clase de investigación son: la comparación de sistemas escolares que requieren que los maestros tengan título universitario (la X) con otros que no exigen esa condición; la comparación de alumnos de cursos que reciben instrucción en lectura veloz con otros que no la reciben; la comparación entre quienes presenciaron determinado programa de televisión y los que no lo hicieron, etc. En marcado contraste con el experimento del diseño 6 «propiamente dicho», que veremos más adelante, no hay en estos casos del diseño 3 ningún medio explícito que permita asegurar que los grupos habrían sido equivalentes de no ser por la X. La ausencia de un medio tal, indicada en el diagrama por las líneas punteadas que separan ambos grupos, señala el próximo factor que requiere control: la *selección.* Si hay diferencias entre O_1 y O_2, ello bien puede deberse al reclutamiento diferencial de las personas que componen los grupos: estos podrían haber diferido aun sin la presencia de X. Como se verá más adelante en el análisis *ex post facto,* la equiparación fundada en características que no sean O suele resultar ineficaz y conducir a error, particularmente en los casos en que las personas que constituyen el «grupo experimental» han procurado la exposición a la X.

Una última variable entremezclada que, por ende, debe incluirse en esta lista es la llamada *mortalidad* experimental, o producción de diferencias $O_1 - O_2$ en grupos, al retirarse en mayor o menor número personas pertenecientes a ellos. Así, aunque en el diseño 3 ambos grupos habían sido alguna vez idénticos, quizá difiriesen ahora, no por haberse producido un cambio en los integrantes individualmente considerados, sino más bien a causa del abandono selectivo de personas de uno de los grupos. En el campo de la investigación educacional, este problema suele encontrarse a menudo en los estudios so-

bre los efectos de la formación universitaria, cuando se comparan las mediciones efectuadas entre alumnos recién ingresados (que no han tenido la X) y los que están a punto de egresar (que la han tenido). Si esos estudios indicaran que las mujeres recién ingresadas son más bellas que las que están por graduarse, rechazaríamos de plano la consecuencia lógica de que nuestro duro curso de capacitación menoscaba la belleza femenina, y señalaríamos en su lugar las dificultades que encuentra una muchacha agraciada para finalizar su carrera antes de contraer matrimonio. Este efecto se clasifica como *mortalidad* experimental. (Por supuesto, si observamos a las *mismas* muchachas cuando acaban de ingresar y cuando egresan, este problema desaparece, con lo cual tenemos el diseño 2.)

4. Tres diseños experimentales propiamente dichos

Los tres diseños fundamentales que vamos a exponer en este capítulo son los recomendados en la actualidad por la literatura metodológica. Son también, como se verá, los más recomendados por nosotros, aun cuando tal respaldo esté sujeto a muchas restricciones concretas en cuanto a la práctica habitual, y dé lugar a que aparezcan algunos signos negativos en el cuadro 1 bajo el título *Validez externa*.

El diseño 4 es el más empleado de los tres; por eso, nos permitiremos la libertad de explayarnos mucho más en su análisis, haciendo de él el centro de convergencia de otras consideraciones, cuya aplicación es más general. Obsérvese que los tres diseños se presentan en forma de comparaciones diversas de una sola X con *ninguna* X. Los diseños que han recibido mayor cantidad de tratamientos por parte de la corriente del experimento factorial de Fisher representan elaboraciones importantes pero tangenciales respecto del hilo conductor de esta obra, y se estudian al final del presente capítulo, a continuación del diseño 6. Ahora bien, esta perspectiva puede servirnos para recordar aquí que comparar X con *no* X es un exceso de simplificación. En realidad la comparación se establece con las actividades específicas desplegadas por el grupo de control durante el período en que el grupo experimental recibe la X. Por lo tanto, sería mejor establecerla entre X_1 y X_c, o entre X_1 y X_o, o entre X_1 y X_2. El que la actividad de esos grupos de control con frecuencia no esté especificada añade un indeseable elemento de ambigüedad a la interpretación del efecto de X.

Teniendo en cuenta todos estos comentarios, continuaremos en este capítulo insistiendo en la convención gráfica de no presentar ninguna X en el grupo de control.

4. Diseño de grupo de control pretest-postest

Controles de validez interna

Algunas de las consideraciones anteriores indujeron a los investigadores psicológicos y educacionales, entre 1900 y 1920, a agregar al diseño 2 un grupo de control, creando el actual diseño ortodoxo con grupo de control. McCall [1923], Solomon [1949] y Boring [1954] fueron en parte los protagonistas de esta historia, y una revisión del *Teachers College Record* de aquel período implica más todavía, pues ya en 1912 se mencionaban grupos de control sin necesidad de mayores explicaciones [p. ej., Pearson, 1912]. Los diseños con grupos de control así introducidos se clasifican en esta sección bajo dos encabezamientos: el presente diseño 4, en el que se emplean grupos equivalentes logrados por aleatorización, y el diseño 10 cuasiexperimental, en el que se utilizan grupos intactos de comparación ya existentes, de equivalencia no asegurada. El diseño 4 adopta la forma

$$\begin{array}{lll} R\,O_1 & X & O_2 \\ R\,O_3 & & O_4 \end{array}$$

Como el diseño controla en forma tan nítida *las siete* hipótesis descritas, las presentaciones que de él se han hecho no han establecido en forma explícita las necesidades de control que satisfacía. En la tradición de las investigaciones del aprendizaje, los efectos prácticos de la *administración de pruebas* parecen ofrecer el primer reconocimiento de la necesidad de contar con un grupo de control. La *maduración* era a menudo el punto crítico de los estudios experimentales en educación, así como del problema naturaleza-cultura (*nature-nurture*) en el campo del desarrollo infantil. En la investigación de los cambios actitudinales, como en los primeros estudios sobre los efectos de las películas cinematográficas, la *historia* puede haber sido la consideración primaria de necesidad. De cualquier manera, creemos conveniente analizar brevemente aquí la forma en que se controlan esos factores, así como las condiciones en que se lo hace.

La *historia* se controla en la medida en que los acontecimientos históricos generales que podrían haber producido una diferencia del tipo $O_1 - O_2$ causarían también una diferencia del tipo $O_3 - O_4$. Adviértase, sin embargo, que muchas supuestas utilizaciones del diseño 4 (o 5, o 6) *no* controlan la existencia de una *historia intrasesional* única. Si a todos los estudiantes, elegidos al azar, que integran el grupo experimental se los trata en una sola sesión, haciéndose lo mismo con los controles, los únicos acontecimientos ocurridos en cada una de esas sesiones y que carecen de importancia (la broma exagerada, el incendio en la otra cuadra, los comentarios introductorios del experimentador, etc.) se convierten en hipótesis rivales que explican la diferencia de $O_1 - O_2$ contra $O_3 - O_4$. Este *no es* un verdadero experimento, aunque se lo presente como paradigma ilustrativo, como en la prueba de Solomon [1949] sobre la enseñanza del alfabeto. (Para ser exactos, tenemos que puntualizar que Solomon lo eligió para ilustrar un aspecto diferente.) Meditando sobre nuestras «mejores prácticas» en relación con ese aspecto, puede que ello carezca de importancia, pero nuestras «mejores prácticas» consisten en presentar experimentos que con harta frecuencia son imposibles de repetir, y esa misma fuente de diferencias «significativas» pero externas bien podría ser una falla importante. Además, en los típicos experimentos que describe el *Journal of Experimental Psychology*, el control de la historia intrasesional, se logra exponiendo a estudiantes y animales a pruebas individuales, y sometiendo aleatoriamente a los estudiantes y los períodos de prueba a condiciones experimentales o de control. Obsérvese, no obstante, que aun con sesiones individuales la historia puede escapar al control si se trabaja con todo el grupo experimental y no con el grupo de control, etc. El diseño 4 requiere que las sesiones experimentales y de control sean simultáneas. Si realizamos sesiones verdaderamente simultáneas, tienen que emplearse distintos experimentadores, y las diferencias entre ellos acaso se conviertan en una forma de historia intrasesional que se confunda con *X*.

La solución óptima es una aleatorización de las sesiones experimentales, aplicando las restricciones requeridas para lograr una representación equilibrada de fuentes de sesgo tan probables como son los experimentadores, la hora, el día de la semana, la parte del semestre, la proximidad de los exámenes, etc. El recurso habitual de trabajar con sujetos experimentales en pequeños grupos —en vez de hacerlo individualmente— es inaceptable si se prescinde de ese agrupamiento en el análisis estadístico. (Cf. más adelante el examen de la asignación de grupos intactos a diversos tratamientos.) Todos los que toman parte en la misma sesión participan de la misma historia in-

trasesional y tienen, por ende, fuentes de similitud distintas de *X*. Si tales sesiones se han asignado al azar, el procedimiento estadístico correcto será el mismo que el que señalamos más adelante para la asignación de aulas intactas a diversos tratamientos. (Para algunos estudios que comprenden la administración de tests en grupos, los distintos tratamientos experimentales pueden distribuirse al azar dentro de un grupo cara a cara, como en el uso de varias formas de un test para estudiar el efecto del orden de dificultad de los ítems. En tales casos, los elementos específicos de la historia intrasesional son comunes a ambos tratamientos y no se convierten en una hipótesis rival aceptable que se confunda con *X* cuando se explican las diferencias obtenidas.)

La *maduración* y la *administración de tests* están controladas en el sentido de que su manifestación en los grupos experimentales y de control debería ser igual. La *instrumentación* se controla con facilidad cuando se dan las condiciones para el control de historia intrasesional, en particular cuando se logra la *O* por medio de reacciones de los estudiantes a un instrumento fijo, como una prueba impresa. Sin embargo, cuando se recurre a observadores o entrevistadores, el problema es ya más grave. Si el número de observadores es suficientemente pequeño para que su asignación a la observación de sesiones individuales no sea aleatoria, no solo habrá que emplear cada observador tanto para las sesiones experimentales como para las de control, sino que además los observadores deberán ignorar cuáles son los estudiantes que reciben cada uno de los distintos tratamientos, a fin de que el conocimiento de ese hecho no sesgue sus puntajes o registros. Tales tendencias al sesgo son causas «confiables» de variancias, como lo confirma la necesidad de contar en las investigaciones médicas con un segundo ciego en la prueba de dos ciegos, y también estudios recientes [Rosenthal, 1959] y anteriores [p. ej., Kennedy y Uphoff, 1939; Stanton y Baker, 1942]. El uso de registros de la interacción grupal, a fin de que los jueces puedan evaluar una serie de secciones aleatorizadas de trascripciones pretest, postest, experimentales y del grupo de control, contribuye al perfecto control de la instrumentación en las investigaciones sobre la conducta escolar y la interacción grupal.

La *regresión* se controla, en lo que a las diferencias de medias concierne y por muy extremo que sea el grupo en los puntajes pretest, si tanto el grupo experimental como el de control se asignan al azar, tomándolos de este mismo conjunto extremo.

En tales casos, el grupo de control regresiona tanto como el experimental. Sin embargo, aun en las condiciones del diseño 4 se producen con frecuencia vacíos interpretativos, a causa de los mecanismos de regresión. Un experimentador puede aprovechar el grupo de control para confirmar los efectos de *X* sobre la media grupal, y después abandonarlo mientras examina cuáles han sido los subgrupos de puntaje pretest del grupo experimental que han registrado mayores influencias. Si todo el grupo acusa una ganancia, llega a la estimulante conclusión artificial de que quienes al principio estaban en la posición más baja han logrado el mayor adelanto, mientras que los que se hallaban en la más elevada quizá no han avanzado lo más mínimo. Este resultado se asegura porque, en condiciones de ganancia media de todo el grupo, el mecanismo de regresión suple el puntaje de ganancia para los participantes con puntaje pretest inferior a la media, y tiende a eliminarlo para quienes en el pretest tenían puntaje elevado. (Si en el conjunto no hubo ningún avance, el experimentador quizá «descubra» por error que aquello se debió a dos efectos mutuamente excluyentes: el avance de los bajos y el retroceso de los altos.) Un modo de evitar esos errores de interpretación es hacer análisis paralelos de aquellos que en el grupo de control presentan puntajes pretest extremos, y fundar las interpretaciones de ganancias diferenciales en comparaciones de los puntajes postest de los correspondientes subgrupos experimentales y de control en el postest. (Nótese, sin embargo, que a causa de las distribuciones asimétricas resultantes de la selección resulta dudosa la conveniencia de las estadísticas de curva normal.)

Se elimina la *selección* como explicación de la diferencia en la medida en que la aleatorización haya asegurado la igualdad grupal en el momento *R*, medida que queda determinada por nuestra estadística de muestreo. Así, la garantía de igualdad es mayor para grandes que para pequeñas cantidades de asignaciones aleatorias. Este supuesto fallará en ocasiones en el grado sugerido por el término de error para la hipótesis de no diferencia. En el diseño 4, ello significa que a veces habrá una aparente diferencia «significativa» entre los puntajes pretest. Por lo tanto, aunque la aleatorización simple o estratificada asegura la asignación no sesgada a los grupos de sujetos experimentales, constituye un medio muy imperfecto para garantizar la equivalencia inicial de dichos grupos. No obstante, es la única forma práctica de hacerlo. Lo decimos así, tan categóricamente, a causa de una muy difundida y errónea pre-

dilección, evidenciada en la investigación educacional durante los últimos treinta años, por la igualación mediante la equiparación. McCall [1923] y Peters y Van Voorhis [1940] contribuyen a perpetuar este equívoco. Como veremos con mayor detalle al estudiar el diseño 10 y el *ex post facto*, la equiparación no constituye una ayuda real cuando se la utiliza para solucionar diferencias iniciales de grupos. Ello no significa que propugnemos la eliminación lisa y llana de este procedimiento como posible aditamento a la aleatorización, como cuando se obtiene mayor precisión estadística asignando estudiantes a pares equiparados y asignando después al azar un miembro de cada par al grupo experimental y otro al de control. En la literatura sobre estadística, esto se designa con el término «bloqueo». Véanse, en particular, los estudios de Cox [1957], Feldt [1958] y Lindquist [1953]. Pero la equiparación como sustituto de la aleatorización es tabú incluso para los diseños cuasiexperimentales que no emplean más que dos grupos naturales intactos, uno experimental y otro de control: aun en ese endeble «experimento» hay medios mejores que la armonización para tratar de corregir diferencias iniciales entre las medias de una y otra muestra.

Los datos de que disponemos gracias al diseño 4 permiten establecer qué *mortalidad* explica aceptablemente la ganancia $O_1 - O_2$. Mortalidad, casos perdidos y casos para los cuales solo se dispone de datos parciales, son difíciles de manejar y por lo común se los trata de disimular. La experimentación típica con métodos educativos se prolonga durante días, semanas o meses. Si se realizan los pretests y postests en las aulas de las que se toman el grupo experimental y el grupo de control, y la condición experimental requiere la concurrencia a determinadas sesiones sin que ocurra lo mismo con la condición de control, la distinta concurrencia a las tres sesiones (pretest, tratamiento y postest) produce una «mortalidad» que puede introducir en la muestra sutiles sesgos. Si de todos los designados en un primer momento como participantes del grupo experimental eliminamos a los que no concurren a las sesiones de prueba, reducimos selectivamente el grupo experimental con un mecanismo que no se aplica en forma similar al grupo de control, sesgando al primero en el sentido de los responsables y sanos. El modo preferido de tratamiento, aunque no de utilización habitual, parece ser el empleo de todos los estudiantes seleccionados, experimentales y de control, que completaron tanto el pretest como el postest, incluso los integrantes del grupo experimental que no obtuvieron la *X*. Es

innegable que este procedimiento atenúa el efecto aparente de *X*, pero evita el sesgo de muestreo, fundándose en el previo supuesto de que no había sesgos de mortalidad más simples. Este supuesto es susceptible de verificación parcial examinando tanto el número como los puntajes pretest de quienes participaron en el pretest pero no en el postest. Es posible que algunas *X* influyeran en esa tasa de abandono, en vez de modificar los puntajes individuales. Por supuesto, aun cuando tales tasas sean las mismas, queda todavía en pie la posibilidad de que se produzcan complicadas interacciones que propenderían a diferenciar el carácter de los abandonos en los grupos experimentales y de control.

El problema de la mortalidad puede observarse con toda claridad en el estudio de *métodos correctivos con voluntarios*. Así, por ejemplo, se invita a un grupo de lectores deficientes de una escuela secundaria a participar en sesiones correctivas voluntarias mientras que otro grupo en las mismas condiciones no es invitado. Del primero de ellos, quizá participen en las sesiones un 30 % de sus integrantes. Los puntajes postests, así como los pretest, provienen de pruebas de lectura estándar administradas a todos los que asistían a clase. No es razonable comparar el 30 % de voluntarios con el total del grupo de control, porque representan a los más preocupados por sus puntajes pretest, los capaces de trabajar con mayor ahínco en su propio mejoramiento, etc.; pero es imposible localizar sus exactos equivalentes en el grupo de control. Aunque tampoco parece justo para la hipótesis de eficacia terapéutica comparar el total del grupo invitado con el total del no invitado, es esta una solución aceptable, si bien moderada. Nótese, empero, que puede ocurrir que lo que produce el efecto sea la invitación misma, y no precisamente la terapia. En general, el grupo de control no invitado debería poseer igual conocimiento de su posición en el pretest que el grupo invitado. Otra posibilidad es invitar a todos los que necesitan sesiones correctivas y asignar los voluntarios a grupos de tratamiento verdaderos y falsos; mas en el estado actual del arte es probable que cualquier terapia suficientemente bien presentada como para que parezca una ayuda al estudiante sea tan eficaz como el tratamiento mismo que se estudia. Cabe, no obstante, destacar la consecuencia innegable de que las pruebas experimentales de la eficacia relativa de dos procedimientos terapéuticos son mucho más fáciles de evaluar que la eficacia absoluta de cualquiera de ellos. La única solución utilizada en la práctica es crear grupos experimentales y de control entre quienes desean

tratamiento correctivo, manipulando los períodos de espera [p. ej., Rogers y Dymond, 1954]. Esto, por supuesto, suscita a veces otras dificultades, como el excesivo abandono por parte del grupo de control con terapia pospuesta. Una aplicación feliz y al parecer no reactiva de un mecanismo de lotería para decidir sobre la aplicación inmediata o diferida, durante un semestre, de un curso correctivo de lectura puede verse en Reed [1956].

Factores que atentan contra la validez externa

Los factores de invalidez externa descritos hasta ahora han sido los que directamente influían sobre los puntajes *O*, y que por sí solos podían producir cambios susceptibles de confundirse con los resultados de *X*, es decir, factores que, una vez agregado el grupo de control, producirían efectos evidentes en este y que se sumarían a los de *X* en el grupo experimental. En el lenguaje propio del análisis de variancias —*historia, maduración, realización de pruebas,* etc.— se los consideró efectos principales, y como tales se los ha controlado en el diseño 4, dándole validez *interna.* Las amenazas a la validez *externa*, por otra parte, pueden considerarse efectos de interacción entre *X* y alguna otra variable. Constituyen, pues, una posible especificidad de los efectos de *X* respecto de algún conjunto de condiciones inconvenientemente limitadas. Como anticipo diremos que, hasta donde nosotros sabemos, en el diseño 4 los efectos observados de *X* pueden ser específicos de grupos cuyo interés fue intensificado por el pretest. Como es natural, no podemos extender nuestras conclusiones al conjunto mayor no sometido a pretest, acerca del cual desearíamos extraer conclusiones.

En este capítulo examinaremos unas cuantas de esas amenazas a la posibilidad de generalización, así como los procedimientos para sortearlas. Es decir que se prefieren estos diseños por razones de validez *externa* o posibilidad de generalización, dado que hay diseños válidos que evitan el pretest y en muchas situaciones (aunque no necesariamente en la investigación educacional) se desea generalizar precisamente con respecto a los grupos no sometidos a pretest. En el campo de la docencia constituyen juicios acerca de la validez *externa* las dudas que con frecuencia se expresan sobre la aplicabilidad práctica de los resultados de ciertos experimentos muy artificiales. La introducción de tales consideraciones en el análisis de los mejores diseños experimentales resulta así atractiva a quien tiene que aplicarlos, pues piensa con toda razón que se las ha soslayado sin motivo en el tratamiento formal común sobre metodología experimental. El análisis siguiente ratificará tales apreciaciones señalando multitud de medios que, sin perder validez *interna*, pueden dar mayor validez externa a los experimentos y bases más firmes de generalización a la práctica docente.

Pero antes de entrar en ese tema nos es inevitable hacer una advertencia que introduce en la ciencia de la inducción unos cuantos problemas espinosos, a causa de la persistente resistencia a aceptar lo que con toda verdad dice Hume, a saber: que *la inducción o generalización nunca tiene una plena justificación lógica.* Al paso que los problemas de validez *interna* son susceptibles de solución dentro de los límites de la lógica de la estadística probabilística, los de validez externa no pueden resolverse en estricto rigor lógico en una forma nítida y concluyente. Generalizar significa siempre mezclar la extrapolación en un cambio no representado en nuestra muestra. Dicha extrapolación se hace *suponiendo* que se conocen las leyes respectivas. Así, si tenemos un diseño 4 internamente válido, el efecto queda demostrado solo a propósito de las condiciones específicas que el grupo experimental y el de control tienen en común, es decir, solo en relación con grupos sometidos a pretest, pero de determinada edad, inteligencia, situación socioeconómica, región geográfica, momento histórico, conjunción estelar, orientación del campo magnético, presión barométrica, nivel de radiaciones gamma, etcétera.

Desde el punto de vista *lógico* no podemos generalizar más allá de dichos límites; es decir que no podemos generalizar en modo alguno. Pero tratamos de hacerlo conjeturando leyes y verificando algunas de dichas generalizaciones en otras condiciones no menos específicas pero diferentes. A lo largo de la historia de cada una de las ciencias, se aprende a «justificar» las generalizaciones propias de ella a causa de la acumulación misma de la experiencia en hacerlo, pero no es esa una generalización lógica, deducible de los resultados del experimento original. En esa situación hacemos, al generalizar, suposiciones sobre leyes aún no demostradas, incluyendo algunas que ni siquiera se indagaron. Así, en la investigación educacional, suponemos por lo común que la orientación del campo magnético no la afecta. Pero sabemos de ciertos estudios que con frecuencia el pretest ha tenido un efecto, y por lo tanto quisiéramos eliminarlo como obstáculo para nuestra generalización. Si hiciéramos una investigación sobre barras de hierro, sabría-

mos por la experiencia que una primera pesada nunca produce efectos reactivos, pero que la orientación del campo magnético, si no se la regulara de manera sistemática, podría limitar gravemente la posibilidad de generalizar nuestros descubrimientos. Los motivos, pues, de invalidación externa son presunciones de leyes generales en la ciencia de una ciencia: conjeturas acerca de los factores que pueden interactuar con nuestras variables de tratamiento según cierta ley, y, por lo tanto, acerca de los que pueden dejarse de lado.

Además de los elementos específicos existe una ley general empírica que nosotros, así como el resto de los hombres de ciencia, aceptamos como supuesto: es la versión moderna de la hipótesis de Mill acerca de la «legalidad» de la naturaleza. Esa versión, menos tajante y drástica, puede enunciarse como el supuesto del «aglutinamiento» (*stickiness*) de la naturaleza: cuanto más cercanos se hallan dos acontecimientos en tiempo, espacio y valor —medido este en cualquiera de sus dimensiones o en todas ellas—, más tienden a ajustarse a las mismas leyes. Si bien las interacciones complejas y las relaciones curvilíneas habrán de confundir presumiblemente los intentos de generalización, tal posibilidad aumenta en relación directa con el grado en que la situación experimental difiere de la situación con respecto a la cual se desea generalizar. Nuestra necesidad de una mayor validez externa será, pues, el requerimiento de la máxima similitud entre experimentos y condiciones de aplicación que sea compatible con la validez interna.

Téngase en cuenta, en este sentido, que las ciencias más «prósperas», como la física y la química, han avanzado sin prestar la menor atención a la representatividad (aunque sí, y mucha, a la reiterabilidad por parte de investigadores independientes). Un laboratorio artificial dentro de una torre de marfil quizá sea una maravilla, pero no será representativo, y a menudo la artificialidad puede resultar imprescindible si se quiere separar analíticamente variables fundamentales para los descubrimientos de muchas ciencias. Pero, sin duda, si no interfiere con la validez interna o el análisis, la validez externa es una consideración de la mayor importancia, sobre todo para una ciencia aplicada como es la pedagogía.

Interacción de las pruebas y X. En estudios del diseño experimental en sí, el peligro que constituye el pretest para la validez externa fue denunciado por primera vez por Solomon [1949], aunque idénticas consideraciones habían llevado antes a algunos experimentadores a la aplicación del diseño 6, que omite el pretest. En especial durante los estudios de cambios de actitud, en que los mismos tests introducen grandes cantidades de contenido extraordinario (p. ej., una tan abundante dosis de declaraciones hostiles como las que se hallan en el test típico de prejuicios), es bastante probable que las actitudes de la persona y su propensión a dejarse persuadir varíen por influjo del pretest. Como psicólogos, dudamos seriamente de la comparabilidad del público que asiste a una proyección de *Gentlemen's Agreement* (un film antiprejuicial) inmediatamente después de habérsele administrado un test de 100 ítems sobre antisemitismo, con otro público que vea la misma película sin que se lo haya sometido precisamente a dicho test. Estas dudas se extienden no solo al efecto principal del pretest, sino también a su efecto sobre la respuesta a la persuasión. Supongamos que esa película en particular fue tan bien realizada que algunas personas llegaron a disfrutarla por su interés romántico, sin darse cuenta siquiera del problema social que planteaba. Tales personas no existirían probablemente en un grupo al que se hubiere administrado un pretest. Si el pretest sensibilizó al público sobre el problema, podría, por medio de una concentración de la atención, intensificar en sí el efecto educativo de *X*. Sería concebible que esa *X* solo resultase eficaz para un grupo al que se hubiese administrado un pretest.

Aunque es frecuente mencionar un efecto sensibilizador de esta índole en comentarios anecdóticos sobre el tema, los pocos resultados publicados de investigaciones indican tanto la ausencia de efectos [p. ej., Anderson, 1959; Duncan y otros, 1957; Glock, 1958; Lana, 1959*a*, 1959*b*; Lana y King, 1960; Piers, 1955; Sobol, 1959; Zeisel, 1947] como un efecto de interacción que equivale a un amortiguador. Así, Solomon [1949] descubrió que administrando un pretest se reducía la eficacia del entrenamiento ortográfico experimental, y Hovland, Lumsdaine y Sheffield [1949] sugirieron que un pretest restringía los efectos persuasivos de las películas cinematográficas. Bien vale la pena evitar este efecto de interacción aunque no sea tan expuesto a error como la sensibilización (ya que los falsos positivos son un problema mayor en nuestra literatura que los falsos negativos, a causa de la gran cantidad de descubrimientos publicados [Campbell, 1959, págs. 168-70]).

Al restringir la validez externa, el efecto del pretest sobre *X* depende, naturalmente, del grado en que tales mediciones repetidas son características del conjunto respecto del cual se

quiere generalizar. En el ámbito de las comunicaciones masivas, la entrevista del investigador y los procedimientos del test de actitud son bastante atípicos. Pero en la investigación pedagógica nos interesa generalizar respecto de una situación en que la administración de tests es una práctica regular. Sobre todo si el experimento puede utilizar como *O* exámenes corrientes tomados en las aulas, pero quizá también si las *O* experimentales son similares a las de utilización habitual, no se produciría ninguna interacción indeseable entre la *administración de los tests* y *X*. Cuando se emplea un test con procedimientos muy poco usuales, o cuando el test implica engaño, reestructuración conceptual o cognitiva, sorpresa, tensión, etc., los diseños con grupos no sometidos a pretest continúan siendo muy convenientes, aunque no imprescindibles.

Interacción entre la selección y X. Aun cuando el diseño 4 controla los efectos de selección a fin de explicar las diferencias entre el grupo experimental y el de control, continúa en pie la posibilidad de que los efectos válidamente demostrados solo se verifiquen en aquella población aislada de la cual se extrajeron a la vez ambos grupos. Esta posibilidad es tanto mayor cuanto más graves son nuestras dificultades de conseguir sujetos para el experimento. Consideremos las posibles consecuencias de un experimento de enseñanza en el cual el investigador se ha visto rechazado por nueve sistemas escolares y aceptado por el décimo. Es casi seguro que ese último difería, en más de un aspecto, de los nueve anteriores, así como del conjunto de escuelas para el que quisiéramos generalizar. Por lo tanto, no es representativo. Podría asegurarse que, en cuanto a la escuela media, su personal tiene más espíritu, menos temor a las inspecciones y más deseo de mejorar. Y aunque los efectos que descubriéramos fuesen internamente válidos, podrían ser específicos de tales escuelas. A fin de poder formular un juicio lo más exacto posible sobre la materia, convendrá que los informes de investigación proporcionen datos sobre cuántas y cómo eran las escuelas y los cursos de los que se solicitó cooperación y la negaron, a fin de que el lector pueda estimar la gravedad de posibles sesgos selectivos. En general, cuanto mayor es la cooperación prestada, mayor el grado en que se afecte la rutina y más elevada nuestra tasa de negativas, mayor será también la oportunidad de que exista un efecto de especificidad de selección.
Aclaremos más puntualmente qué es lo que en realidad significa la «*interacción* entre selección y *X*». Si estuviésemos

por realizar un estudio dentro de una única escuela voluntaria, empleando la asignación aleatoria de sujetos a grupos experimentales y de control, no nos preocuparía el «efecto principal» de la escuela en sí. Si este factor elevara por igual la media del grupo experimental y la del de control, no se causaría daño alguno. Pero si existiesen en la escuela características que hicieran más eficaz al tratamiento experimental en ella que en la población de escuelas que constituyen el verdadero objetivo de la prueba, las consecuencias podrían ser graves. Queremos estar seguros de que puede menospreciarse la interacción entre las características de la escuela (probablemente relacionadas con el hecho de que es voluntaria) y los tratamientos experimentales aplicados. Algunas variables experimentales podrían ser bastante sensibles a las características de la escuela, lo cual quiere decir que interactuarían con ellas; otras, no. La interacción *podría* darse en escuelas con CI medios similares, o no presentarse allí donde las diferencias de CI fuesen elevadas. Sería de esperar, sin embargo, una mayor probabilidad de interacción si las escuelas difiriesen mucho en distintas características que si fuesen análogas.
A menudo se producen importantes sesgos de muestreo a causa de la inercia de los experimentadores, que no conceden a una selección más representativa de escuelas la oportunidad de negarse a participar. De ahí que la mayoría de las investigaciones sobre educación se hagan en los establecimientos que cuentan con mayor porcentaje de alumnos hijos de profesores universitarios. Aunque es imposible la representatividad perfecta en el muestreo, y aun se la descuida casi en absoluto en muchas ciencias (por ejemplo, en la mayoría de los estudios publicados en el *Journal of Experimental Psychology*), puede y debe aspirarse a ella como a un desiderátum en la investigación educacional. Una forma de aumentarla es reducir el número de alumnos o aulas participantes que pertenezcan a un colegio o nivel dado y aumentar la cantidad de escuelas y niveles en que se lleve a cabo el experimento. Es obvio que nunca vamos a realizar experimentos sobre muestras que representen a todas las aulas de Estados Unidos o del mundo. Solo poco a poco aprenderemos hasta dónde se puede generalizar un descubrimiento internamente válido, por medio de comprobaciones empíricas en ese sentido. Pero tales intentos de generalización tendrán éxito más a menudo si en el experimento original se demuestra el fenómeno en una amplia variedad de condiciones.
En cuanto a los signos positivos y negativos que aparecen en

el cuadro 1, resulta evidente que nada seguro puede consignarse en esa columna. Se la presenta, no obstante, porque los requisitos de algunos diseños exageran o atenúan la gravedad de este problema. El diseño 4, dentro del ámbito de las actitudes sociales, es tan exigente en lo que a cooperación por parte de los participantes se refiere, que en definitiva la investigación solo se hace con un público cautivo en vez de realizarla con ciudadanos comunes, que son a quienes quisiéramos referirnos. En una situación de esa índole, el diseño 4 merecería un signo negativo en cuanto a selección. No obstante, en la investigación pedagógica nuestro universo de interés está constituido por un público cautivo para el cual se pueden obtener diseños 4 de elevada representatividad.

Otras interacciones con X. De manera parecida, las interacciones de *X* con los demás factores pueden examinarse como amenazas a la validez *externa.* La *mortalidad* diferencial sería un producto de *X* y no una interacción con ella. La interacción de la *instrumentación* con *X* se ha incluido implícitamente en el análisis de validez *interna*, ya que un efecto específico de instrumentación ante la presencia de *X* falsearía el verdadero efecto de *X* (p. ej., cuando los observadores asignan puntajes, conocen las hipótesis y saben cuáles son los estudiantes que recibieron *X*). Una amenaza a la validez externa es la posibilidad de que los efectos sean específicos de los instrumentos particulares (tests, observadores, medidores, etc.) empleados en el estudio. Si en todos los tratamientos se utilizan observadores o entrevistadores múltiples, tales interacciones pueden estudiarse directamente [Stanley, 1961*a*]. La *regresión* no interacciona con *X*.

La *maduración* tiene consecuencias de especificidad de selección: los resultados pueden ser específicos de un determinado grupo etario, del cansancio, etc. La interacción de la *historia* y *X* implicaría que el efecto había sido específico de las condiciones históricas del experimento, y aunque su observación es válida, no se lo hallaría en otras.

El hecho de que el experimento se llevase a cabo en el trascurso de una guerra, o a continuación de haber fracasado una huelga de maestros, etc., podría producir una reacción frente a *X* que no aparecería en otras circunstancias. Si tuviésemos que preparar un modelo de muestreo para este problema, nos gustaría que el experimento se repitiese en una muestra aleatoria de ocasiones pretéritas y futuras, lo cual, como es obvio, resulta imposible. Además, compartimos con

otras ciencias el supuesto empírico de que no existen leyes que dependan en verdad del tiempo, que los efectos de la *historia*, cuando los haya, se deberán a combinaciones específicas de condiciones de estímulo que se dieron en aquel momento, y que llegarán a incorporarse en definitiva a leyes generales independientes del tiempo [Neyman, 1960]. (Tal vez parezca que las cosmologías de un «universo en expansión» requieren una restricción de esta afirmación, pero no en formas pertinentes a lo que ahora estudiamos.) Sin embargo, la feliz reiteración de los resultados de la investigación en distintos tiempos y situaciones aumenta nuestra confianza en el valor de la generalización, al disminuir la probabilidad de la interacción con la *historia.*

Estos distintos factores no se han incluido como otros tantos encabezamientos de columnas en el cuadro 1, porque no ofrecen bases firmes de discriminación entre diferentes diseños.

Dispositivos reactivos. En el experimento psicológico común, si no en la investigación educativa, la obvia artificialidad de la situación experimental y la conciencia del estudiante de que está participando en un experimento son causas más que suficientes de carencia de representatividad. Para sujetos humanos, se proyecta una tarea de resolución de problemas de orden más elevado, en la cual se reacciona contra los procedimientos y el tratamiento experimental no solo en razón de sus simples valores de estímulo, sino también por su función de claves para interpretar la intención del experimentador. El representar cargos, el adivinar la intención, el prepararse para la inspección, el sentir cada cual que «soy un conejillo de Indias», o muchas otras actitudes así generadas, no son en modo alguno representativas de la verdadera situación escolar; parecen calificar más bien el efecto de *X*, dificultando gravemente la generalización. Cuando es imposible evitar tales *dispositivos reactivos*, habría que continuar de cualquier manera con los experimentos de esa índole que tengan validez interna, pero resulta obvia la conveniencia de evitarlos cuando ello sea posible. Al hacer esta afirmación adherimos en parte a la conocida crítica antiexperimental que es frecuente en los consejos de educación y entre los docentes, contra la «futilidad» de «toda esa experimentación». Nuestra más moderada conclusión no es, sin embargo, que habría que abandonar la investigación por ese motivo, sino más bien que, a causa de él, habría que mejorarla. A este respecto tenemos unas cuantas sugerencias que ofrecer.

Cualquier aspecto del procedimiento experimental puede producir ese resultado de *dispositivos reactivos*. La administración de pretests, prescindiendo de su contenido, puede hacerlo, y parte de la interacción del *pretest* con *X* puede ser de ese tipo, aunque hay poderosas razones para sospechar de los aspectos mismos de contenido de la aplicación del test. El sistema de aleatorización y asignación a tratamientos quizá sea de esa índole. Consideremos el efecto que se produce sobre una clase cuando (como en Solomon [1949] se hace pasar a una habitación separada a la mitad de los alumnos, elegidos al azar. Ese acto, más la presencia de «maestros» extraños, tiene que crear por fuerza expectaciones de hechos desusados, suscitándose así el asombro y una activa curiosidad en cuanto a su objeto y finalidad. La presentación del tratamiento *X*, si fuese un acontecimiento inusitado, podría tener un efecto similar. Es de presumir que aun el postest, en un diseño 6 de postest solamente, podría crear esas mismas actitudes. Cuanto más evidente sea la conexión entre el tratamiento experimental y el contenido postest, más probable será ese efecto.

En el campo de los cambios de opinión pública, esos dispositivos reactivos suelen ser difíciles de evitar. Pero en la mayor parte de la investigación de métodos educativos no hay necesidad de que los estudiantes sepan que se está realizando un experimento. (Sería muy conveniente que también los maestros lo ignorasen, a la manera del doble ciego en medicina, pero por lo común esto suele ser imposible.) Varios recursos permiten disimularlo. Si las *X* son variables sobre acontecimientos usuales en el aula, pero que se producen a intervalos bastante largos dentro del calendario escolar, un tercio de la batalla se habrá ganado si los tratamientos mencionados se aplican sin previo anuncio. En forma similar, si se incluyen las *O* en exámenes regulares, se llena el segundo requisito. Si las *X* son comunicaciones centradas en determinados estudiantes, puede lograrse la aleatorización sin necesidad de trasportar físicamente muestras aleatorias equivalentes a aulas distintas, etcétera.

A la luz de estas consideraciones, así como de observaciones personales de los experimentadores que han publicado datos pese a tener un *rapport* tan pobre que sus hallazgos eran bastante engañosos, los autores del presente volumen van llegando poco a poco a la conclusión de que la experimentación dentro de las escuelas debe realizarse, siempre que sea posible, con el personal regular de ella, en especial cuando los

descubrimientos hayan de generalizarse a otras situaciones escolares.

En estos momentos, parecen estar en boga dos tipos principales de «experimentación» dentro de las escuelas: 1) estudios «impuestos» a la escuela por alguien de fuera, que persigue sus propios intereses y cuyo objetivo no es que la escuela emprenda una acción inmediata (o cambio), y 2) el llamado investigador «de acción», que procura que sean los maestros mismos quienes «experimenten», tomado este término en sentido muy amplio. En el primer caso los resultados pueden ser rigurosos pero no aplicables. En el segundo, en cambio, quizá sean muy aplicables pero probablemente no son «ciertos», a causa de una gran falta de rigor en la investigación. Otro modelo posible es que las ideas para la investigación escolar partan de los maestros y directores, se elaboren los diseños para someterlas a prueba en cooperación con especialistas en metodología de investigación y luego se encarguen de la mayor parte de la experimentación los promotores de la idea. Los análisis estadísticos respectivos podría realizarlos el investigador metodologista, y los resultados los volvería a introducir al grupo un intermediario idóneo (supervisor, director de investigaciones del consejo escolar, etc.) que hubiera servido en tal carácter durante todo el proceso. De esa manera se lograrían resultados pertinentes y «correctos». La forma de realizar investigación *básica* con un sistema de esta índole es un problema en gran parte sin resolver aún, pero los estudios podrían ser cada vez menos *ad hoc* y orientarse más hacia los aspectos teóricos, bajo la supervisión de un intermediario competente.

Aunque no tenemos en esta obra la intención de destacar los buenos o malos ejemplos observables en la literatura especializada, un reciente estudio de Page [1958] indica una utilización tan buena de estos aspectos (evitando dispositivos reactivos, logrando representatividad de muestreo y evitando las interacciones entre las pruebas y *X*), que vamos a citarlos aquí como ilustración concreta de la práctica óptima. Su estudio indica que breves comentarios escritos agregados a exámenes objetivos que se devuelven a los alumnos hacen mejorar el rendimiento en pruebas objetivas posteriores. A esta conclusión se llegó actuando con 74 maestros, 12 consejos escolares, 6 niveles o grados (7-12), 5 niveles de rendimiento (A, B, C, D, F) y gran variedad de sujetos; no hubo casi prueba alguna de efectos de interacción.

Los alumnos y las clases se eligieron al azar. Se empleó como

pretest el primer examen objetivo regular en cada clase. Arrojando un dado especial, el maestro asignaba alumnos a grupos de tratamiento y, según los casos, agregaba o no comentarios escritos a la prueba. La siguiente prueba objetiva, tomada de acuerdo con la programación normal, pasó a ser el postest. Hasta donde pudo determinarse, ninguno de los 2.139 alumnos se enteró de la experimentación. Pocos son los procedimientos de instrucciones que se prestan a esta tan disimulada aleatorización, ya que por lo común la comunicación oral necesaria se dirige a toda la clase y no a algunos individuos. (Las comunicaciones escritas permiten la aleatorización, aunque la captación, por parte del estudiante, de los distintos tratamientos constituye un problema.) Teniendo en cuenta estos ideales los investigadores pueden lograr que los experimentos tengan menos características reactivas que en la actualidad.

Por medio de exámenes regulares tomados en el aula, o tests presentados como exámenes regulares y análogos en su contenido, a la vez que mediante procedimientos alternativos de enseñanza presentados, sin previo aviso ni petición de disculpas, en el curso de las actividades escolares, es probable que en la mayoría de los casos puedan evitarse estas dos causas de dispositivos reactivos. A veces, en grandes escuelas secundarias o en universidades donde los alumnos se inscriben en cursos populares dictados en determinados horarios y después se los asigna en forma arbitraria a múltiples divisiones simultáneas, podrían lograrse secciones de equivalencia aleatoria mediante el control del proceso de asignación (véase en Siegel y Siegel [1957] la aplicación de un proceso aleatorio natural que se aprovechó en esta forma). Sin embargo, por la acción de historias intragrupales únicas, tales secciones, al principio equivalentes, se tornan con el correr del tiempo en segmentos cada vez más diferenciados.

La solución a este problema, aplicable en general, es trasladar la aleatorización al aula tomada como unidad y construir grupos experimentales y de control, constituido cada uno de ellos por muchas aulas asignadas al azar [véase Lindquist, 1940, 1953]. Por lo común, aunque no es imprescindible, los cursos se clasificarían para su análisis sobre la base de factores como escuela, maestro o (cuando este tenga varias clases), hora, asignatura, nivel intelectual medio, etc.; de ellos se asignarían por un proceso aleatorio varios grupos de tratamiento experimental. Ya se han realizado algunos estudios de esta índole, pero creemos que pronto se generalizarán. Nótese que el test de significación apropiado *no consiste* en mezclar todos

los estudiantes como si se los hubiese asignado al azar. Los detalles se estudiarán en el capítulo siguiente.

Tests de significación para el diseño 4

Hay que distinguir el diseño experimental del uso de tests estadísticos de significación. El primero es el arte de lograr comparaciones interpretables y, como tal, sería necesario aunque el producto final consistiera en porcentajes graficados, fotografías de grupos en acción, etc. En todos estos casos, la interpretabilidad de los «resultados» depende del control sobre los factores a que hemos hecho referencia. Si la comparación es interpretable, se requieren tests estadísticos de significación para decidir si las diferencias obtenidas exceden o no las fluctuaciones previsibles cuando no existan verdaderas diferencias para muestras de ese tamaño. El uso de tests de significación presume que es factible establecer comparaciones entre los grupos, y que la diferencia descubierta es interpretable, pero no da pruebas de ello. De ahí que nos gustaría exponer el diseño experimental sobre la base del sentido común y de consideraciones no matemáticas. Esperamos que la mayor parte de esta obra resulte accesible a los estudiantes de ciencias de la educación que carezcan todavía de preparación estadística. No obstante, hay que reconocer que la cuestión de los procedimientos estadísticos está íntimamente vinculada al diseño experimental, razón por la cual ofrecemos estos comentarios particulares sobre el tema. [Véase, asimismo, Green y Tukey, 1960; Kaiser, 1960; Nunnally, 1960, y Rozeboom, 1960.]

Una estadística errónea de uso común. Aunque el diseño 4 es el común y frecuente, los tests de significación que con él se utilizan son a menudo erróneos, incompletos o inapropiados. Al aplicar la «razón crítica» común o prueba t a ese diseño experimental estándar, muchos investigadores han computado dos t: una para la diferencia pretest-postest en el grupo experimental y otra para la ganancia pretest-postest en el grupo de control. Si la primera resulta «estadísticamente significativa» y la otra «no», llegan a la conclusión de que X tuvo un efecto, sin ninguna comparación estadística directa entre el grupo experimental y el de control. A menudo las condiciones fueron tales que, de haberse aplicado una prueba más apropiada, la diferencia no habría sido significativa (como cuando los valores de significación son casos límites y el gru-

po de control indica una ganancia que casi alcanza el nivel de significación). Windle [1954] y Cantor [1956] han demostrado la frecuencia de este error.

Utilización de puntajes de ganancia y covariancia. La prueba aceptable de uso más común consiste en computar para cada grupo puntajes de ganancia pretest-postest y calcular una *t* entre los grupos experimentales y de control sobre la base de esos puntajes. El «bloqueo» o «nivelación» aleatoria de puntajes pretest y el análisis de covariancia utilizando como covariable los puntajes de pretest son, por lo común, preferibles a las simples comparaciones de puntajes de ganancia. Puesto que la mayor parte de los experimentos en educación no acusan diferencias significativas, y por lo tanto no suelen informarse, el uso de este análisis más preciso parece ser muy conveniente. Considerando la labor que implica conducir un experimento, el trabajo de realizar el análisis correcto es relativamente trivial. Para más detalles, pueden consultarse tratamientos estándar de análisis del tipo Fisher [véanse también Cox, 1957, 1958; Feldt, 1958, y Lindquist, 1953].

Aspectos estadísticos de la asignación aleatoria a tratamientos de cursos intactos. La estadística habitual solo resulta apropiada en casos de asignación aleatoria de alumnos individuales a los tratamientos. Si, en cambio, se asignaran cursos intactos, las fórmulas precedentes darían un término de error demasiado pequeño, pues, como es natural, el procedimiento de aleatorización habrá sido más «global» y se habrán utilizado menos acontecimientos aleatorios. Lindquist [1953, págs. 172-89] ha suministrado el fundamento lógico y las fórmulas para la realización de un correcto análisis. En esencia, se emplean las medias de la clase como observaciones básicas, y se prueban los efectos del tratamiento contra variaciones en esas medias. Un análisis de covariancia utilizaría como covariable medias pretest.

Aspectos estadísticos de la validez interna. Las observaciones precedentes se hicieron a fin de dar a conocer la ortodoxia estadística relativa al diseño experimental. Las siguientes representan un esfuerzo por ampliar o corregir esa ortodoxia, extendiendo al terreno de la estadística del muestreo una inferencia de la distinción entre *validez externa* y *validez interna.* Los principios estadísticos antes analizados implican en su totalidad el muestreo en un universo infinitamente grande, más apropiado para una encuesta de opinión pública que

para el experimento habitual de laboratorio. En casos muy raros, como el estudio de Page [1958], hay un muestreo real tomado de un gran universo predesignado, que se apropia las fórmulas habituales. En el extremo opuesto se encuentra el experimento de laboratorio presentado en el *Journal of Experimental Psychology*, por ejemplo, en el que la *validez interna* ha sido la única consideración y *todos* los integrantes de un pequeño universo único se asignaron a los grupos de tratamiento. En este tipo de prueba se pone gran énfasis en el procedimiento aleatorio, pero no a fin de asegurarse la representatividad respecto de otra población mayor, sino al exclusivo efecto de igualar los grupos experimentales y de control o los distintos grupos de tratamiento. La aleatorización se aplica, pues, a una población finita muy reducida, que es en realidad la suma de los grupos experimentales y de control.

Esta posición extrema sobre el universo de muestreo se justifica cuando se describen procedimientos de laboratorio de esta índole: se solicitan voluntarios, prometiéndoles o no una gratificación en dinero, puntajes de personalidad, puntajes para la aprobación de cursos, o cumplimiento de un requisito obligatorio que de todos modos tendrán que satisfacer en algún momento del curso académico. A medida que llegan, se los va asignando al azar a los distintos tratamientos. Cuando se ha alcanzado determinado número de sujetos, se interrumpe el experimento. Ni siquiera ha habido una selección aleatoria entre los integrantes de una lista mucho mayor de voluntarios. Los primeros constituyen una muestra sesgada y el universo total «muestreado» cambia de un día a otro a medida que el experimento continúa, que se requiere más presión para reclutar voluntarios, etc. En un momento dado se detiene el procedimiento, después de haberse utilizado a todos los miembros designables del universo en uno u otro de los grupos de tratamiento. Nótese que los sesgos implicados de muestreo no amenazan en lo más mínimo la equivalencia aleatoria de los grupos de tratamiento, sino solo su «representatividad».

Consideremos ahora a un científico más meticuloso, que de una clase integrada por 250 personas extrae 100 al azar, se pone en contacto con ellas por carta o por teléfono y, después de entrevistarlos, los asigna, también al azar, a grupos de tratamiento. Por supuesto, unos 20 de ellos no pueden ajustarse al horario de laboratorio, están enfermos, etc., por lo cual se ha producido una redefinición implícita del universo. Y aunque gracias a su perseverancia consiga los 100, lo que ha-

brá ganado, desde el punto de vista de la representatividad, será la posibilidad de generalizar con seguridad estadística a propósito del curso del año 1961 de Psicología Educacional A en la Escuela Normal del Estado. Este nuevo universo, aunque mayor, carece de positivo interés científico. Sus límites no son los estatuidos por ninguna teoría científica. Los aspectos de verdadero interés para la generalización deberán explorarse por medio del muestreo de experimentos realizados en otros lugares. Por supuesto, al ser menos seleccionados sus alumnos, se tiene una mayor validez externa, pero no ganancia suficiente para que la mayoría de los psicólogos experimentales consideren que se compensa con ello el esfuerzo realizado.

Resulta, en general, obvio que el fin principal que se persigue con la aleatorización en experimentos de laboratorio es la validez interna, no la externa. Por tanto, habría que utilizar márgenes de error más reducidos y apropiados, basados en pequeños universos finitos. Siguiendo a Kempthorne [1955] y Wilk y Kempthorne [1956], creemos que el modelo correcto es la aleatorización en urnas en vez de la extracción de muestras de un universo. De ese modo se dispone de un test no paramétrico más apropiado y preciso, en el cual se toman los puntajes obtenidos en los grupos experimentales y de control y se los asigna una y otra vez a dos «urnas», generando empírica o matemáticamente una distribución de diferencias medias que resultan en su totalidad de asignaciones aleatorias de esos puntajes particulares. Tal distribución constituye el criterio con que debería compararse la diferencia media obtenida. Cuando exista una «interacción posición-tratamiento» (heterogeneidad de efectos reales entre los sujetos), esa distribución tendrá una variabilidad menor que la correspondiente distribución adoptada en la prueba común. Con estos comentarios no pretendemos modificar mucho la actual práctica en la administración de tests de significación en la investigación pedagógica. Las soluciones exactas son difíciles de conseguir y, por lo común, muy laboriosas. La aleatorización por urnas, por ejemplo, suele exigir la utilización de computadoras de gran velocidad. La dirección del error es conocida: el empleo de la estadística tradicional es demasiado conservador, con una excesiva tendencia a decir «no se registran efectos». Si juzgamos que nuestras publicaciones están saturadas de «falsos positivos», es decir, de información sobre efectos que no resiste la prueba de una validación cruzada (como acaece, por cierto, con la psicología experimental

y social, aunque no todavía con la investigación pedagógica), ese error —si lo es— será siempre preferible. La posibilidad de subestimar la significación es mayor cuando solo hay dos condiciones experimentales y se emplean todos los sujetos disponibles [Wilk y Kempthorne, 1955, pág. 1154].

5. Diseño de cuatro grupos de Solomon

Aunque el diseño 4 se usa más, el 5, denominado diseño de cuatro grupos de Solomon [1949] tiene con razón un mayor prestigio y constituye la primera consideración explícita de factores de *validez externa*. El diseño es el siguiente:

R	O_1	X	O_2
R	O_3		O_4
R		X	O_5
R			O_6

Trazando en forma paralela los elementos del diseño 4 (O_1 a O_4) con los grupos experimental y de control sin pretest, cabe determinar tanto los efectos principales de la *realización de la prueba* como la interacción entre ella y X. De ese modo, no solo se aumenta la posibilidad de generalizar, sino que además se repite el efecto de X en cuatro formas diferentes: $O_2 > O_1$, $O_2 > O_4$, $O_5 > O_6$ y $O_5 > O_3$. Las inestabilidades concretas de la experimentación son tales que, si esas comparaciones concuerdan, el vigor de la inferencia queda muy incrementado. Otra contribución indirecta a la posibilidad de generalizar los hallazgos experimentales es también que, en virtud de la experiencia con el diseño 5 en cualquier ámbito de investigación dado, se averigua la posibilidad general de interacciones de «pruebas por X», pudiéndose así interpretar mejor los diseños 4, tanto futuros como pasados. Asimismo, puede advertirse (comparando O_6 con O_1 y O_3) un efecto combinado de maduración e historia.

Pruebas estadísticas para el diseño 5

No hay ningún procedimiento estadístico particular que utilice a un mismo tiempo los seis conjuntos de observaciones. Las asimetrías del diseño descartan el análisis de la variancia

de puntajes. (Las sugerencias de Solomon a este respecto se consideran inaceptables.) Dejando de lado los pretests, salvo como un nuevo «tratamiento» coordinado con X, se pueden estudiar los puntajes postest mediante un simple análisis 2×2 del diseño de variancia:

	Sin X	Con X
Con administración de pretest	O_4	O_2
Sin administración de pretest	O_6	O_5

Sobre la base de las medias de las columnas se estima el efecto principal de X; de las medias de las filas, el efecto principal del pretest y de las medias de los casilleros, la interacción entre la aplicación del test y X. Si los efectos principales e interactivos de la aplicación de las pruebas son muy pequeños, acaso sea conveniente realizar un análisis de covariancia de O_4 contra O_2, con los puntajes del pretest por covariable.

6. Diseño de grupo de control con postest únicamente

El pretest es un concepto muy arraigado en el pensamiento de los investigadores en los campos de la educación y la psicología, pero en realidad no es imprescindible para los diseños experimentales propiamente dichos. Por razones psicológicas, es difícil renunciar a «tener la seguridad» de que los grupos experimentales y de control eran «iguales» antes del tratamiento experimental diferencial. No obstante, la aleatorización implica la mayor seguridad, aplicable a cualquier fin, de la carencia de sesgos iniciales entre grupos. Dentro de los márgenes de confianza establecidos por las pruebas de significación, la aleatorización puede ser suficiente, sin necesidad de recurrir al pretest. En realidad, casi todos los experimentos agrícolas realizados en la tradición de Fisher [1925, 1935] carecen de pretest. Más todavía, en investigación pedagógica, sobre todo en los grados primarios, tenemos que experimentar a menudo con métodos que permitan la introducción inicial de elementos absolutamente nuevos, para los cuales son imposibles los pretests en el sentido ordinario del término, lo mismo que estarían fuera de lugar los referidos a la presunta culpabilidad o inocencia en un estudio acerca de los efectos

de la información presentada al jurado por el abogado defensor. El diseño 6 responde a esa necesidad, y además es apropiado para todas las situaciones en que podrían utilizarse los diseños 4 o 5, es decir, aquellas en que es posible una verdadera aleatorización. Su forma es la siguiente:

R	X	O_1
R		O_2

Si bien este diseño se utilizaba ya en la década de 1920, la mayoría de los textos metodológicos no lo han recomendado. Ello se debió en parte a que se lo confundía con el diseño 3, y también a la falta de confianza en la aleatorización como procedimiento de igualación.

Puede considerarse que este diseño comprende los últimos dos grupos del diseño de cuatro grupos de Solomon; controla la aplicación del test como efecto principal y la interacción, pero, a diferencia del diseño 5, no los mide. Sin embargo, esa medición es tangencial a la cuestión básica de si X tuvo o no un efecto. Así, pues, el diseño 5 es preferible al 6 por las razones apuntadas, pero las mayores ventajas del 5 quizá no justifiquen el esfuerzo que demanda (más del doble). Asimismo, el diseño 6 es por lo común preferible al 4, a menos que haya alguna duda a propósito de la autenticidad del proceso aleatorio de asignación. El diseño 6 se usa demasiado poco en investigación educacional y psicológica.

Pero en el caso de repetición de pruebas, que se presenta con frecuencia en la investigación educacional, si se dispone de antecedentes apropiados en materia de variables, se los debería emplear para bloqueo o nivelación, o como covariables. Esta recomendación la hacemos por dos motivos. Primero, porque las pruebas estadísticas en que se apoya el diseño 4 son más decisivas que las existentes para el 6. El esfuerzo que exige el diseño 4 anula esta ventaja en la mayor parte de las situaciones de investigación, pero no ocurriría así si se dispusiese en forma automática de antecedentes apropiados sobre puntajes. En segundo lugar, la disponibilidad de puntajes pretest permite examinar la interacción de X y el nivel de habilidad en el pretest, explorando así más a fondo la posibilidad de generalizar el hallazgo. Algo similar puede hacerse a propósito del diseño 6, empleando otras medidas disponibles en vez del pretest, pero estas consideraciones, sumadas al hecho de que para la investigación pedagógica los tests frecuentes son característicos del universo al cual se quieren

extender las generalizaciones, pueden invertir el criterio de preferir por lo común el diseño 6 al 4. Nótese asimismo que para cualquier mortalidad sustancial entre *R* y el postest los datos de pretest del diseño 4 ofrecen mayores oportunidades de eliminar la hipótesis de mortalidad diferencial entre los grupos experimental y de control.

Aun así, hay muchos problemas para los cuales no se dispone de pretests, o estos resultan inconvenientes o capaces de provocar reacciones, y para esos casos es preciso seguir insistiendo, en muchos sectores, acerca de la legitimidad del diseño 6. Además de los estudios sobre el modo de enseñar material nuevo, queda una gran cantidad de casos en los que la *X* y la *O* postest pueden entregarse a los alumnos o grupos como un solo «paquete» natural, y un pretest resultaría molesto. Tales situaciones se producen con frecuencia en los mismos procedimientos de prueba, así como en estudios de instrucciones distintas, planillas de respuesta de formato diferente, etc. Algo similar ocurre con los estudios sobre campañas para reclutar voluntarios, etc. En los casos en que hay que guardar el anonimato del alumno, el diseño 6 suele ser el más conveniente, encarándose entonces la aleatorización por medio del ordenamiento mezclado de materiales destinados a la distribución.

Aspectos estadísticos del diseño 6

El modo más sencillo sería la prueba *t*. El diseño 6 es quizá la única situación para la cual esa prueba es óptima. Sin embargo, se pueden emplear el análisis de covariancia y el bloqueo de «variables sujeto» [Underwood, 1957*b*], así como niveles anteriores de educación, puntajes en tests, ocupación de los padres, etc., consiguiéndose así mayor poder del test de significación, muy similar al que brinda un pretest. No es necesario que el pretest y el postest sean idénticos. A menudo serán formas diferentes «del mismo» test y por lo tanto menos idénticos que una repetición del pretest. La mayor precisión obtenida se vincula en forma directa con el grado de covariancia, y aunque esta suele ser más elevada en formas alternadas «del mismo» test que en tests «diferentes», se trata de una cuestión de grado tan confiable y factorialmente compleja como la superioridad eventual de un promedio puntual respecto de un breve «pretest». Adviértase, sin embargo, que un promedio puntual no es por lo común conveniente como medición postest, a causa de su probable insensibilidad a *X* si se lo compara con una medición más específicamente apropiada en contenido y oportunidad. No tiene mucha importancia decidir si ese seudodiseño de pretest debe clasificarse como 6 o como 4. Tendría las ventajas del primero, ya que evitaría una sesión pretest introducida por el experimentador, así como la «reveladora» repetición de un contenido poco usual idéntico o muy similar (como en los estudios de cambios de actitud). Por estas razones la inclusión del diseño 6 bajo el título de «Dispositivos reactivos» debería ser algo más positiva que respecto de los diseños 4 y 5. La justificación de esta diferencia es, por cierto, mucho más válida para las ciencias sociales en general que para la investigación sobre instrucción pedagógica.

Diseños factoriales

Sobre la base conceptual de los tres diseños anteriores, pero en particular el 4 y el 6, pueden ampliarse las complejas elaboraciones típicas de los diseños factoriales de Fisher, agregando otros grupos con otras *X*. En un criterio típico de clasificación única o análisis de la variancia «en un solo sentido», tendríamos varios «niveles» del tratamiento, por ejemplo, X_1, X_2, X_3, etc., y quizá también un grupo X_0 (ausencia de *X*). Si se considera el grupo de control como uno de los tratamientos, habría en los diseños 4 y 6 un grupo para cada tratamiento. En el diseño 5 habría dos grupos (uno sometido a pretest, el otro no) para cada tratamiento, y aun sería posible un análisis de variancia de doble clasificación («en dos sentidos»). No tenemos noticia de que se hayan realizado diseños del tipo 5 en más de dos niveles. Por lo común, si nos preocupa la interacción pretest, empleamos el diseño 6, a causa del gran número de grupos que de no hacerlo así serían necesarios. Muy a menudo se utilizarán dos o más variables de tratamiento, una en cada uno de los distintos «niveles», dando una serie de grupos que podrían designarse X_{a1} X_{b1}, X_{a1} X_{b2}, X_{a1} X_{b3}, . . . , X_{a2} X_{b1}, etcétera.

Tales elaboraciones, complicadas con intentos de economizar eliminando algunas de las posibles permutaciones de X_a por X_b, han producido parte de los inquietantes misterios del diseño factorial (bloques aleatorizados, parcelas divididas, cuadrados grecolatinos, repetición fraccional, confusión, etc.),

origen de la enorme brecha que separa las metodologías avanzadas de las tradicionales en el ámbito de la investigación educacional. Esperamos que esta obra ayude a salvar ese vacío por medio de una continuidad con la metodología tradicional y las consideraciones dictadas por el sentido común que el estudiante lleva siempre consigo. También estimamos que gran parte de lo que debe enseñarse sobre diseño experimental se entiende mejor si se lo expone en forma de diseños de dos tratamientos, sin interferencia de otras complicaciones. No obstante, la exposición completa de los problemas planteados por el uso común provocará una comprensión mayor tanto de la necesidad como de la localización de modernos enfoques. Al buscar la forma más eficaz de resumir el anticuado pero tan difundido diseño 4 nos vimos ya constreñidos a disponer de un análisis de covariancia, casi no utilizado en esta situación. Y en el diseño 5, con un problema de dos tratamientos que se elabora sólo para obtener controles necesarios, nos alejamos de las relaciones críticas o pruebas t, y nos introducimos en la estadística del análisis de variancia.

Los detalles de los análisis estadísticos para diseños factoriales no pueden enseñarse ni aun esbozarse siquiera en esta obra. Edwards [1960], Ferguson [1959], Johnson y Jackson [1959] y Lindquist [1953] presentan a los investigadores pedagógicos aspectos elementales de tales métodos. Confiamos, sin embargo, en que las explicaciones siguientes permitirán alguna mayor comprensión de ciertas alternativas y complejidades de particular relevancia en los aspectos de diseño analizados en nuestra obra. Las complejidades que tenemos que analizar no comprenden las razones comunes para recurrir a cuadrados latinos ni a muchos otros diseños incompletos en que el conocimiento de ciertas interacciones se sacrifica por meras razones de costo. (Pero el uso de cuadrados latinos como sustituto de los grupos de control en los casos en que no hay modo de aleatorizar se estudiará más adelante, como diseño cuasiexperimental 11.) La razón de haber prescindido aquí de esos diseños incompletos es que para el problema de validez externa resulta muy conveniente contar con un conocimiento detallado de las interacciones, sobre todo en una ciencia que ha tenido problemas para repetir los descubrimientos de un investigador en otro ambiente distinto [véase Wilk y Kempthorne, 1957]. Los conceptos que tratamos de exponer en este capítulo son los de la interacción, las clasificaciones inclusivas y las clasificaciones cruzadas, y los modelos factoriales finitos, fijos, aleatorios y mixtos.

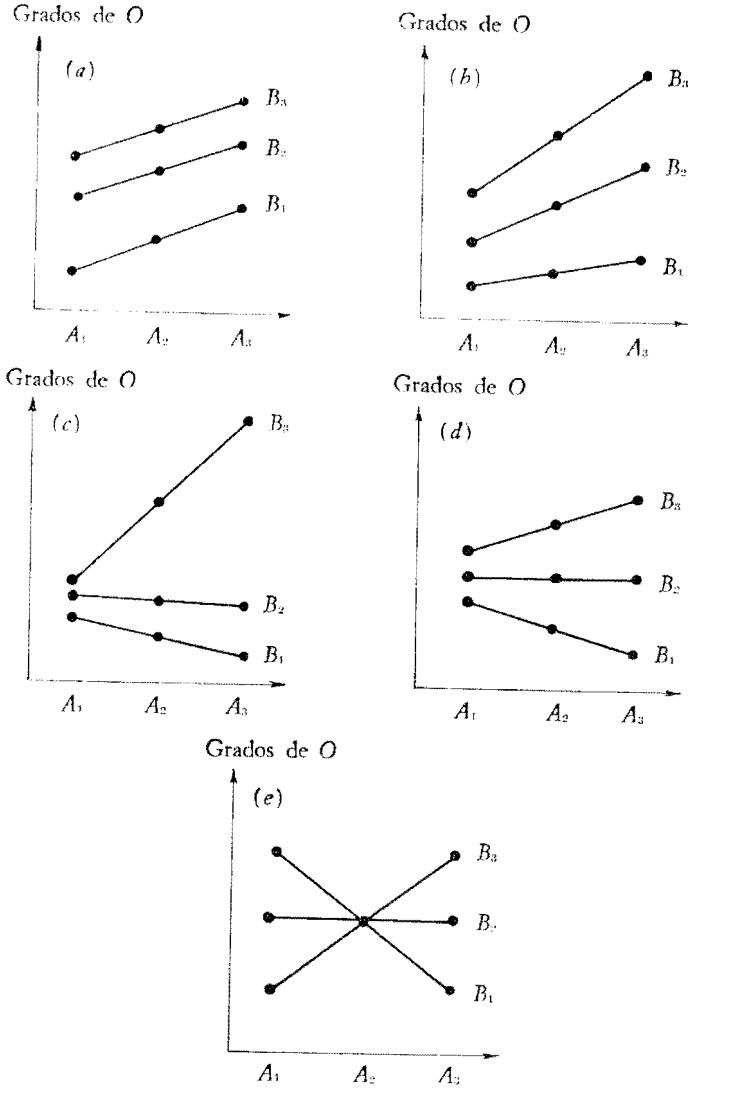
Interacción

Ya hemos utilizado este concepto en situaciones en que, según creemos, el lector no especializado lo habrá encontrado comprensible. Como antes, insistimos aquí en las consecuencias relativas a la posibilidad de generalizar. Expondremos en forma gráfica (figura 2) cinco posibles resultados de un diseño con tres niveles, compuesto cada uno de ellos de X_a y X_b, que denominaremos A y B. (Puesto que se han de graficar tres dimensiones [A, B y O] en dos, son varias las presentaciones que resultan posibles, de las cuales no emplearemos más que una.) En la figura 2(a) puede apreciarse un notable efecto principal tanto para A como para B, pero ninguna interacción. (Hay, por supuesto, una suma de efectos —siendo A_3, B_3 el más fuerte—, pero ninguna interacción, ya que los efectos son aditivos.) En todos los demás casos encontramos interacciones significativas adicionales, o bien en reemplazo de los efectos principales de A y B. Es decir que la ley sobre el efecto de A varía de acuerdo con el valor específico de B.

En este sentido, los efectos de interacción son reglas de especificidad de efecto y, por tanto, conducen al intento de generalización. El efecto de interacción en 2(d) es, sin duda alguna, de ese orden. Aquí, A no produce un efecto principal (es decir que, si se promedian los valores de las tres B para cada A, resulta una línea horizontal). Pero cuando se mantiene B en el nivel 1, los aumentos en A producen un efecto decreciente, en tanto que si se mantiene B en el nivel 3, A tiene un efecto incremental. Nótese que si el experimentador sólo hubiese variado A, manteniendo B constante en el nivel 1, los resultados, aunque internamente válidos, hubiesen dado pie a que se hiciesen generalizaciones erróneas a propósito de B_2 y B_3. La característica de múltiple factorial del diseño ha llevado, pues, a realizar valiosas exploraciones sobre la posible generalización o validez externa de cualquier enunciación sumaria sobre el efecto principal de A. Las limitaciones de la posibilidad de generalizar o de la espeficidad de los efectos aparecen en el análisis estadístico como interacciones significativas.

La figura 2(e) representa una forma de interacción aún más extrema: ni A ni B producen efecto principal alguno (no resultan reglas generales sobre qué nivel de ambas es mejor) pero las interacciones son fuertes y bien definidas. Consideremos un resultado hipotético de esta índole. Supongamos

Figura 2. *Algunos resultados posibles de un diseño factorial* 3×3.



que tres clases de maestros son, en general, de la misma eficiencia (p. ej., los improvisadores espontáneos, los que preparan a conciencia su trabajo y los que supervisan con esmero la tarea de sus alumnos). Asimismo, tres métodos de enseñanza resultan, en general, de idéntica eficacia (p. ej., discusión en grupo, exposición formal y supervisión individual). En tal caso, aun en ausencia de «efectos principales» en cualquiera de los tipos de maestros o métodos pedagógicos, podría ocurrir que estos segundos tuviesen gran interacción con la modalidad del maestro: el improvisador espontáneo tendría más éxito con la discusión en grupo y menos con la supervisión individual, mientras que el acostumbrado a seguir de cerca a sus alumnos alcanzaría los mejores resultados en la supervisión individual y los peores en el sistema de la discusión en grupo.

Desde este punto de vista, cabe distinguir los tipos de interacciones significativas halladas. Quizá nos resulte provechoso un concepto como el de «interacciones monótonas». Nótese que en $2(b)$, como en el $2(a)$, hay un efecto principal tanto de A como de B, y que A produce el mismo efecto direccional en cualquier panel separado de valores de B. En consecuencia, nos sentimos mucho más seguros si generalizamos a situaciones nuevas la expectativa de aumento en O con aumentos en A que si lo hacemos en $2(c)$, que podría producir también efectos principales significativos en A y B, así como una interacción A-B significativa. En realidad, podríamos estar casi tan seguros de la generalidad del efecto principal de A en el caso $2(b)$ como en el $2(a)$, libre este de interacción. Por cierto que al interpretar efectos con miras a la generalización se las debería graficar y examinar bien, en todos sus detalles. Algunas interacciones «monótonas» o unidireccionales producen pocas limitaciones —y a veces ninguna— sobre la especificidad. (Véase en Lubin [1961] un profundo estudio de este problema.)

Clasificaciones inclusivas

En los ejemplos dados hasta aquí, todos los criterios de clasificación (las A y las B) se han «cruzado» con todos los demás criterios. Es decir que todos los niveles de A se han dado con todos los niveles de B. Sin embargo, el análisis de variancia no se limita a esa situación.

Hasta ahora hemos utilizado, a título de ilustración, criterios

de clasificación que eran «tratamientos experimentales». Otros tipos de criterios de clasificación, como el sexo y la edad de los alumnos, podrían introducirse en muchos experimentos en forma de clasificaciones plenamente cruzadas. Pero a fin de incorporar los usos más comunes de clasificaciones «inclusivas», presentaremos la posibilidad de criterios de clasificación menos obvios. Uno de ellos es «maestros». Operando en el nivel de cruzado total, se podría hacer una prueba en una escuela secundaria en la cual diez maestros emplearían uno de los dos métodos posibles para enseñar una determinada asignatura a distintos cursos experimentales. En ese caso los maestros serían un criterio de clasificación absolutamente cruzado, pues cada uno de ellos constituiría un «nivel» diferente. El «efecto principal» de los «maestros» sería la evidencia de que algunos de ellos son mejores que otros, con prescindencia del método que emplearon. (Los estudiantes o las clases se habrán asignado al azar; de lo contrario se confundirían las idiosincrasias del maestro con las diferencias de selección.) Una interacción significativa entre maestros y métodos significaría que el método que mejor funcionó había dependido del docente en particular a quien se estaba considerando.

Supongamos ahora, siguiendo una interacción de esta índole, que nos interesa conocer si, en general, una técnica dada es mejor para maestros que para maestras. Si dividimos ahora nuestros diez maestros en cinco hombres y cinco mujeres, se obtiene una clasificación «inclusiva» en el sentido de que la clasificación maestros, aunque todavía útil, no considera los sexos; es decir que el mismo maestro no aparece en uno y otro sexo, en tanto que cada maestro y cada sexo sí considera los métodos. Esta inclusividad exige un análisis algo distinto de aquel en que todas las clasificaciones se cruzan entre sí. (Un análisis ilustrativo puede verse en Green y Tukey [1960] y Stanley [1961*a*].) Además, quedan eliminadas ciertas interacciones de las variables inclusivas. Así, no son computables —ni tienen, en realidad, sentido desde el punto de vista conceptual— las interacciones maestros-sexo y maestros-sexo-método.

«Maestros» también podría convertirse en una clasificación inclusiva, si el experimento anterior se extendiese a varias escuelas, de manera que ellas viniesen a constituir un criterio de clasificación (en el cual los efectos principales acusarían diferencias en la tasa de aprendizaje de los alumnos de los distintos establecimientos). En tal caso los maestros serían por

lo común «inclusivos» dentro de las escuelas, ya que lo habitual es que un maestro dé clases en un solo colegio. En este caso es concebible una interacción maestro-escuela, pero no se la podría computar a menos que todos los maestros enseñasen en uno y otro establecimiento, en cuyo caso maestros y escuelas serían «cruzados», no «inclusivos».

A los alumnos, o sujetos de un experimento, también se los puede tratar como criterio de clasificación. En un mecanismo totalmente cruzado, a cada alumno se lo somete a cada uno de los distintos tratamientos, pero en muchos casos entra en varios, aunque no en todos, los tratamientos; es decir que no se produce el fenómeno de la inclusión. Un caso frecuente es el estudio, durante el aprendizaje, de datos obtenidos por pruebas individuales. Aquí podríamos tener curvas de aprendizaje para los distintos alumnos, divididos estos entre dos métodos de estudio. Podrían analizarse las interacciones prueba-método y alumno-prueba, pero no alumno-método. Si a los alumnos se los clasifica por el sexo, se genera también un fenómeno de inclusión.

Casi todas las variables de interés en la experimentación educacional pueden cruzarse con otras variables y no tienen por qué ser objeto de inclusión. Excepciones notables, además de las ya mencionadas, son la edad cronológica, la edad mental, el grado escolar (primero, segundo, etc.) y el nivel socioeconómico. El lector inteligente habrá notado que las variables independientes, o los criterios de clasificación, pertenecen a distintos tipos: 1) variables manipuladas, como el método de enseñanza, que el experimentador puede asignar a voluntad; 2) aspectos potencialmente manipulables, como materias estudiadas, que el experimentador podría asignar de alguna manera aleatoria entre los alumnos que utiliza, pero que rara vez lo hace; 3) aspectos relativamente fijos del ambiente, como comunidad, escuela o nivel socioeconómico, fuera del control directo del experimentador pero que sirven de bases explícitas para la estratificación de la prueba; 4) características «orgánicas» de los alumnos, como edad, estatura, peso y sexo, y 5) características de reacción de los alumnos, como puntajes en distintos tests. Por lo común las variables independientes manipuladas de la clase 1 son de interés fundamental, mientras que las variables independientes no mamanipuladas de las clases 3, 4 y a veces 5 sirven para aumentar la precisión y revelar hasta qué punto son generalizables los efectos de las variables manipuladas. Las variables de clase 5 aparecen de ordinario como covariables o variables depen-

dientes. Otra forma de considerar las variables independientes es como intrínsecamente ordenadas (grado, nivel socioeconómico, estatura, pruebas, etc.) o no ordenadas (método de enseñanza, asignatura, maestro, sexo, etc.). A menudo, los efectos de las variables ordenadas suelen analizarse más a fondo, a fin de ver si la tendencia es lineal, cuadrática, cúbica o de grado más elevado [Grant, 1956; Myers, 1959].

Modelos finitos, aleatorios, fijos y mixtos

Hace poco, estimulados por el trabajo inédito de Tukey del año 1949, varios estadísticos matemáticos crearon modelos «finitos» para el análisis de variancias que aplican al muestreo de «niveles» de factores experimentales (variables independientes) los principios, bien elaborados ya, del muestreo en poblaciones finitas. Scheffé [1956] publicó una reseña histórica de aquel desarrollo clarificador. Se dispone de medias cuadráticas esperadas, que ayudan a determinar «términos de error» apropiados [Stanley, 1956] para el diseño factorial totalmente aleatorizado de tres clasificaciones. Los modelos finitos resultan de particular provecho porque pueden generalizarse con facilidad a situaciones en que uno o más de los factores son aleatorios o fijos. Ferguson dio una sencilla explicación de aquellas extensiones en 1959.

En vez de presentar fórmulas, recurriremos a una ilustración verbal para mostrar cómo difieren entre sí las selecciones finita, aleatoria y fija de niveles de un factor. Supongamos que en un experimento dado los «maestros» constituyen una de las distintas bases de clasificación (es decir, variables independientes). Si se dispone de 50 maestros, se podrían extraer 5 de ellos *al azar* y utilizarlos en el estudio. Aparecería entonces en algunas de nuestras fórmulas un coeficiente de muestreo de factores ($1 - 5/50$) o 0,9. Si se utilizara el total de 50 maestros, constituirían un efecto «fijo», y el coeficiente se convertiría en ($1 - 50/50$) $= 0$. Por lo contrario, si existiese una población prácticamente infinita de maestros, 50 de ellos elegidos al azar constituirían un porcentaje infinitesimal, por lo que en cada efecto «aleatorio» el coeficiente tendería a 1. Los anteriores coeficientes modifican las fórmulas de medias cuadráticas esperadas, y por lo tanto de términos de «error». Más detalles pueden verse en Brownlee [1960], Cornfield y Tukey [1956], Ferguson [1959], Wilk y Kempthorne [1956] y Winer [1962].

Otras dimensiones de extensión

Antes de abandonar los «verdaderos» experimentos a propósito de los diseños cuasiexperimentales, queremos explorar algunas otras extensiones desde este simple núcleo, aplicables a todos los diseños que se verán más adelante.

Aplicación de tests en busca de efectos mediatos

En la esfera de la persuasión —bastante afín a la de la educación y la enseñanza—, Hovland y sus colegas comprobaron, en reiteradas oportunidades, que los efectos a largo plazo son no solo cuantitativa sino también cualitativamente diferentes. Estos efectos son mayores que los inmediatos en las actitudes generales, aunque más débiles en algunas actitudes específicas [Hovland, Lumsdaine y Sheffield, 1949]. Las afirmaciones de una persona desacreditada carecen de efecto persuasivo inmediato, pero ese efecto puede resultar significativo un mes más adelante, a menos que se recuerde a los interlocutores de qué fuente provienen [Hovland, Janis y Kelley, 1953]. Estos descubrimientos nos alertan contra la práctica de establecer toda nuestra evaluación experimental de los métodos pedagógicos sobre la base de postests o mediciones inmediatas realizadas en cualquier punto aislado del tiempo.

A pesar de los problemas incomparablemente mayores de ejecución implicados (y la incomodidad que ello constituye para el desarrollo del programa de nueve meses de una tesis de doctorado), nos permitimos recomendar que en la planificación de las investigaciones se incluyan períodos de postests de un mes, seis meses y un año.

Cuando las mediciones del postest consistan en calificaciones y puntajes de exámenes que de todos modos van a obtenerse, ese estudio será un simple problema de contabilidad (y mortalidad). Pero cuando sea el experimentador quien introduzca las *O*, casi todos los autores consideran que la repetición de mediciones postest con los mismos alumnos sería más engañosa que el pretest. Así se ha comprobado por cierto en investigaciones sobre memoria [p. ej., Underwood, 1957*a*]. Al paso que el grupo de Hovland recurría a la típica aplicación de un pretest (diseño 4), ellos organizaron grupos separados experimentales y de control para cada aplazamiento cronológico del postest, por ejemplo:

R	*O*	*X*	*O*	
R	*O*		*O*	
R	*O*	*X*		*O*
R	*O*			*O*

Para los diseños 5 o 6 se exigiría una duplicación similar de grupos. Nótese que este diseño carece de control perfecto para su propósito de comparar las diferencias en los efectos como función del tiempo trascurrido, puesto que tales diferencias podrían deberse también a la interacción entre *X* y los acontecimientos históricos específicos que se produjeron entre la aplicación de los postests de corto y de largo plazo. Un control completo de esta posibilidad lleva a la elaboración de diseños más complejos todavía. A causa de los grandes gastos que esos estudios exigen, salvo cuando las *O* se obtienen por algún mecanismo rutinario, parece recomendable que quienes realizan estudios empleando *O* institucionalizadas reiteradamente disponibles aprovechen la ventaja que ello representa y realicen observaciones ulteriores de los efectos en varios momentos sucesivos.

Generalización a otras X: Variabilidad en la ejecución de X

El objetivo de la ciencia comprende la generalización, no solo a otras poblaciones y momentos cronológicos, sino también a representaciones distintas del mismo tratamiento, es decir, a otras representaciones que en teoría deberían ser idénticas, pero que no lo son en determinados aspectos que, en principio, carecen de importancia. Esta meta es contraria a la demanda de un mayor control experimental, que a menudo resulta evidente y que conduce al deseo de obtener en cada repetición una réplica *exacta* de *X*. Así, al estudiar el efecto de una apelación emocional frente a otra racional, y volviendo al ejemplo del individuo que hace declaraciones públicas, podríamos conseguir que la misma persona se dirigiese a los distintos tipos de grupo empleando todos los grados de persuasión posibles o, con mayor rigor todavía, grabar sus declaraciones a fin de que todos los públicos incluidos en un determinado tratamiento oyesen «exactamente el mismo» mensaje. Aparentemente, esto sería mejor que si varias personas hablasen una sola vez cada una en los distintos niveles de persuasión, ya que en este caso «no sabríamos con exactitud» qué estímulos experimentales se aplicaron en cada sesión

Pero ocurre lo contrario si por «saber» interpretamos la habilidad para seleccionar la correcta clasificación abstracta del tratamiento y trasmitir eficazmente la información a nuevos destinatarios. Con la entrevista grabada hemos repetido cada vez muchos aspectos específicos carentes de importancia; hasta donde nos fue dado conocer, el efecto pudo haberse creado por esos detalles y no por las características que incluimos adrede. No obstante, si tenemos muchos ejemplos independientes, los detalles específicos sin importancia no serán susceptibles de repetición en cada caso, y por tanto será más probable que nuestra interpretación de la causa de los efectos sea correcta.

Consideremos, por ejemplo, la comparación de Guetzkow, Kelly y McKeachie [1954] entre los métodos de enseñanza por disertación y por discusión. Nuestro «conocimiento» de cuáles fueron los tratamientos experimentales, en el sentido de poder extraer recomendaciones para otros maestros, es mejor *porque* se emplearon ocho docentes, cada uno de los cuales interpretó cada método a su manera, en vez de utilizar uno solo, o de hacer que los ocho memorizasen detalles comunes no incluidos en la descripción abstracta de los procedimientos comparados. (Como en Guetzkow y otros [1954], esa ejecución heterogénea de *X* debería complementarse, de ser posible, con la práctica de que cada tratamiento lo ejecutara cada uno de los participantes en el experimento, para que ningún elemento específico sin importancia se confundiera con un tratamiento específico. A fin de poder estimar la significación de la interacción maestro-método cuando se emplean cursos intactos, convendría que cada maestro aplicara dos veces cada método.)

En un ejemplo más sencillo, un estudio del efecto del sexo del docente sobre los primeros pasos de instrucción aritmética debería utilizar no uno solo, sino muchos ejemplos de cada sexo. Aunque esta es una precaución obvia, no siempre se la ha respetado, como lo señala Hammond [1954]. El problema constituye un aspecto de la insistencia de Brunswik [1956] en el diseño representativo. Underwood [1957*b*, págs. 281-87] ha sostenido, sobre fundamentos similares, una posición contraria a la estandarización o réplica exacta de los aparatos utilizados en los distintos estudios, de manera compatible con su vigoroso operacionalismo.

Generalización a otras X: Refinamiento secuencial de X y grupos de control noveles

En cualquier experimento la *X* real es un complicado conjunto de lo que eventualmente se habrá de conceptualizar como distintas variables. Una vez detectado un efecto fuerte y definido, el curso del proceso científico exige que se realicen nuevos experimentos que refinen la *X*, destacando bien los aspectos más esenciales al efecto. Ese refinamiento se logrará por medio de tratamientos definidos y presentados en forma más particular y concreta, o bien organizando nuevos grupos de control, que igualen al grupo experimental en un número cada vez mayor de aspectos del tratamiento, reduciendo las diferencias a características más específicas de la compleja *X* original. El grupo de control falso y el de control con operación simulada que se utilizan en la investigación médica son ejemplos de ello. Los experimentos anteriores demostraron un efecto internamente válido, pero que, no obstante, pudo haberse debido a que el paciente sabía que se lo sometía al tratamiento, o bien al shock quirúrgico, y no a las propiedades específicas de la droga o a la remoción del tejido cerebral: de ahí la introducción de los controles especiales para prever esas posibilidades. La generalización a otras *X* es un proceso exploratorio de extrapolaciones sugeridas por la teoría, pero sujetas a la experiencia, en cuyo transcurso es posible que el mencionado refinamiento de *X* represente un importante papel.

Generalización a otras O

Así como una *X* dada arrastra un bagaje de caracteres específicos teóricamente sin importancia, pero que pueden resultar los causantes del efecto, así también cualquier *O* dada, cualquier instrumento de medición, es un complejo en el cual el contenido correspondiente está necesariamente inserto en una situación instrumental concreta, cuyos detalles son marginales a la finalidad teórica. Así, cuando utilizamos lápices y planillas de respuesta con calificación mecánica IBM solemos hacerlo por razones de conveniencia y no porque queramos incluir en nuestros puntajes la variancia debida a la habilidad de los empleados, la familiaridad con el formulario del test, la exactitud en la observancia de las instrucciones, etc. Asimismo, nuestro examen de la competencia específica en un

tema objeto de investigación por medio de pruebas consistentes en la redacción de ensayos habrá de efectuarse empleando como vehículos la habilidad literaria y el uso del vocabulario y, por lo tanto, deberá contener la variancia debida a esas fuentes que, con frecuencia, no son importantes para nuestros fines. Dada esa complejidad inherente a cualquier *O*, nos encontramos con un problema cuando queremos generalizar los resultados a otras *O* posibles. ¿A qué aspecto de nuestra *O* experimental se debió aquel efecto internamente válido? Como la finalidad de la enseñanza no es solo la de preparar individuos para futuros exámenes de ensayo y objetivos, debe tomarse siempre en cuenta ese problema de la validez externa o la posibilidad de generalización.

Una vez más, desde el punto de vista conceptual, la solución no está en confiar a ciegas en que se tendrán mediciones «puras» sin complejidades carentes de importancia, sino más bien en utilizar medidas múltiples en las cuales los medios y detalles específicos sin importancia sean todo lo diferentes que sea posible, al paso que el contenido común que nos preocupa esté presente en todos y cada uno de ellos. Dentro de un experimento aislado, es más lo que puede hacerse en este sentido por las *O* que por las *X*, pues en un solo experimento se pueden lograr muchas mediciones de efecto (es decir, variables dependientes). En el estudio de Guetzkow, Kelly y McKeachie [1954], se notaron efectos no solo en los exámenes regulares de curso y en pruebas especiales de actitud introducidas a este fin, sino también en comportamientos ulteriores, como la elección de carrera y la inscripción en cursos superiores sobre el mismo tema. (Aquellos comportamientos resultaron de igual sensibilidad a las diferencias de tratamiento que las mediciones del test.) *Las O múltiples deberían ser un requisito ortodoxo en cualquier estudio sobre métodos de enseñanza.* En el plano más simple, deberían aplicarse tanto exámenes objetivos como de ensayo [véanse Stanley y Beeman, 1956], junto con índices de participación en clase, etc. (Una extensión de esta perspectiva a la cuestión de la validez de los tests se hallará en Campbell y Fiske [1959] y Campbell [1960].)

5. Diseños cuasiexperimentales[1]

Son muchas las situaciones sociales en que el investigador puede introducir algo similar al diseño experimental en su programación de procedimientos para la recopilación de datos (p. ej., el *cuándo* y el *a quién* de la medición), aunque carezca de control total acerca de la programación de estímulos experimentales (el *cuándo* y el *a quién* de la exposición y la capacidad de aleatorizarla), que permite realizar un auténtico experimento. En general, tales situaciones pueden considerarse como diseños cuasiexperimentales. Uno de los propósitos de esta obra es inducir a que se utilicen estos cuasiexperimentos y se aumente el conocimiento de los tipos de situaciones en que se dan oportunidades para su empleo. Pero precisamente porque se carece de control experimental total, *es* imprescindible que el investigador tenga un conocimiento a fondo de cuáles son las variables específicas que su diseño particular no controla. Por esa necesidad de evaluar cuasiexperimentos, más que para satisfacer la de comprender los experimentos propiamente dichos, se prepararon las listas de verificación de fuentes de invalidación en los cuadros 1, 2 y 3.

El estudiante o posible investigador medio que haya leído el capítulo anterior quizá se encuentre con más problemas sin resolver en el diseño de un experimento que los que había considerado al comienzo que pudieran plantearse siquiera. Será para su bien si todo ello lo induce al diseño y ejecución de mejores experimentos y a una mayor circunspección al extraer conclusiones de los resultados obtenidos. Constituirá, sin embargo, un efecto secundario indeseable si crea en él la sensación de desesperanza en cuanto al logro del control experimental y lo induce a abandonar tales esfuerzos para

1 Este capítulo recurre en su mayor parte a D. T. Campbell, «Diseños cuasiexperimentales para su aplicación en situaciones sociales naturales» en D. T. Campbell, *Experimenting, validating, knowing: problems of method in the social sciences*, Nueva York: McGraw-Hill, en preparación.

acogerse a la práctica de métodos de investigación más informales todavía. Además, esta larga lista de fuentes de invalidación podría, con mayor probabilidad aún, reducir la voluntad de realizar los diseños cuasiexperimentales en que se advierta desde un primer momento que se carece de pleno control experimental. Este resultado sería la antítesis de lo que nos habíamos propuesto.

Desde el punto de vista de su interpretación definitiva y del intento de adaptarlo al proceso evolutivo de la ciencia, todo experimento es imperfecto. Lo que puede lograr una lista de verificación de criterios de validez es que el experimentador tenga más conciencia de las imperfecciones residuales que implica su diseño, para poder determinar en los puntos pertinentes las distintas interpretaciones de sus datos. Por supuesto que debería diseñar el mejor experimento que la situación permitiera, y buscar con el mayor empeño los laboratorios artificiales y naturales que ofrecieran las mejores oportunidades de control. Pero, además de todo ello, tendría que seguir experimentando e interpretando con plena conciencia de los puntos donde los resultados son aún equívocos. Esa conciencia es importante en los experimentos en que se ha ejercitado un control «total», pero es imprescindible en los diseños cuasiexperimentales.

En persecución de ese objetivo general, reseñaremos a esta altura de nuestra obra las ventajas e inconvenientes de un conjunto heterogéneo de diseños cuasiexperimentales, cada uno de los cuales merece utilizarse *allí donde no haya otros mejores susceptibles de que se los aplique*. Veremos primero tres diseños experimentales unigrupales. Después, cinco tipos generales de experimentos multigrupales. Una sección aparte se ocupará de la correlación, los diseños *ex post facto*, los estudios en panel y otros temas análogos.

Algunos comentarios preliminares sobre la teoría de la experimentación

Este capítulo está destinado en principio al experimentador que desee sacar sus investigaciones del laboratorio para trasladarlas a la situación operativa. Sin embargo, los autores no pueden dejar de reconocer que los psicólogos experimentales quizá verán con suspicacia cualquier intento de recomendación de estudios en que el control experimental no sea com-

pleto. En parte para justificar el presente trabajo ante esos monitores, ofrecemos algunos comentarios generales acerca de la función de los experimentos en la ciencia, con la convicción de que son compatibles con la mayor parte de las modernas teorías científicas que ellos fundan en la perspectiva de una posible psicología general de los procesos inductivos [Campbell, 1959].

La ciencia, como otros procesos cognitivos, comprende la formulación de teorías, hipótesis, modelos, etc., así como la aceptación o el rechazo de ellos en virtud de algún conjunto de criterios externos. La experimentación pertenece a esa segunda fase, la del desbrozamiento, el rechazo y la revisión. Podemos suponer para nuestra ciencia una ecología en la cual el número de posibles hipótesis positivas exceda en mucho al de las hipótesis que a la larga demostrarán ser compatibles con nuestras observaciones. *La característica predominante de la tarea de compilación de datos para la prueba de teorías es, pues, el rechazo de hipótesis inadecuadas.* Para conseguirlo resulta provechoso cualquier ordenamiento de observaciones en virtud del cual se desautorice la teoría correspondiente, incluyendo diseños cuasiexperimentales de menor eficacia que los verdaderos experimentos.

Cabe preguntarse, sin embargo, si tales diseños imperfectos no vendrán a confirmar con falsedad una teoría inadecuada, descarriando del buen camino los siguientes esfuerzos y desperdiciando el espacio de nuestras publicaciones con las docenas de estudios que parecen necesitarse para desarraigar un falso positivo de notable divulgación. Es este un grave riesgo, que, no obstante, debemos encarar, y del cual participan —en esencia, ya que no en grado— los «verdaderos» experimentos de los diseños 4, 5 y 6. En un sentido muy fundamental, los resultados experimentales nunca «confirman» ni «demuestran» una teoría: más bien, la teoría triunfante está probada y escapa a la refutación. La palabra «demostrar», a menudo empleada para designar la validez deductiva, ha adquirido en nuestra generación un significado impropio, tanto respecto de sus anteriores aplicaciones como a su utilización actual en procedimientos inductivos, como la experimentación científica. Los resultados de un experimento «ponen a prueba» pero no «prueban» una teoría. Una hipótesis bien fundada es aquella que ha sobrevivido en reiteradas ocasiones a esos exámenes, pero que siempre puede ser desplazada por otra nueva investigación.

En la actualidad se entiende que la «hipótesis nula», utilizada

a menudo por conveniencia al enunciar la hipótesis de un experimento, nunca puede ser «aceptada» en virtud de los datos obtenidos; solo cabe «rechazarla» o «no rechazarla». De igual modo, las hipótesis más generales de hecho nunca se «confirman»; cuando por conveniencia utilizamos ese término queremos significar, más bien, que la hipótesis fue expuesta a refutación y salió airosa de ella. Este punto de vista es compatible con todas las filosofías humanas de la ciencia que proclaman la imposibilidad de obtener pruebas concluyentes para leyes inductivas. En trabajos recientes, Hanson [1958] y Popper [1959] han sido taxativos a este respecto. Muchos conjuntos de datos recopilados en la investigación educacional tienen poco o ningún valor indagatorio, y muchos grupos de hipótesis son tan intrincados que no se los puede confirmar por medio de los mecanismos de sondeo disponibles. No deseamos en modo alguno acrecentar la aceptabilidad de esa seudoinvestigación. Creemos que los diseños de investigación que estudiamos más adelante son, sin embargo, lo bastante indagatorios para merecer que se los utilice allí *donde no se disponga de otros medios de estudio más eficaces.*

Aunque correcta, la idea de que los experimentos jamás «confirman» la teoría contradice de tal forma nuestras actitudes y experiencias como científicos que nos resulta casi intolerable. En particular, ese énfasis parece poco aceptable frente a las ruidosas y llamativas confirmaciones obtenidas en física y química, donde los trabajos de experimentación pueden ajustarse con minuciosidad, sobre muchos puntos de medición, a una compleja curva prevista por la teoría. Y para la mayoría de nosotros la perspectiva se torna inaceptable, en sentido fenomenológico, cuando se la extiende a las conclusiones inductivas de la visión. Resulta, por ejemplo, difícil comprender que las mesas y sillas que «vemos» ante nosotros no sean «confirmadas» o «aprobadas» por la evidencia visual, sino que consistan en «meras» hipótesis sobre objetos externos aún no desautorizadas por las múltiples indagaciones del sistema óptico. Hay algo de razón en ese rechazo.

Se confiere a una teoría diversos grados de «confirmación» a tenor de la mayor o menor cantidad de *hipótesis rivales aceptables* de que se dispone para explicar la información. Cuanto menos hipótesis rivales queden, mayor será el grado de «confirmación». Es de presumir que en cualquier etapa de la recopilación de datos, aun para la más avanzada de las ciencias, hay muchas teorías compatibles con la información, en

especial si se consideran todas las teorías que abarcan circunstancias complejas. Sin embargo, en la práctica se dispone de pocas teorías —cuando las hay— que hagan frente a las «bien establecidas» o a las que han sido verificadas a fondo mediante complicados experimentos; tampoco se proponen seriamente esas teorías rivales. Dicha escasez es el equivalente epistemológico de la afirmación positiva de la teoría que parecen ofrecer los experimentos espectaculares. Una escasez semejante de hipótesis rivales se da en el conocimiento fenoménicamente positivo que por contraste parece ofrecer, por ejemplo, la visión a la comparativa ambigüedad de la exploración táctil a ciegas.

Dentro de esta perspectiva, la lista de fuentes de invalidación que controlan los diseños experimentales puede considerarse como una enumeración de hipótesis —a menudo aceptables— rivales de la hipótesis de que la variable experimental ha surtido un efecto. Donde un diseño experimental «controla» uno de esos factores, se limita a hacer insostenible esta hipótesis rival, aun cuando, en virtud tal vez de complicadas coincidencias, continúe operando para producir el resultado experimental. Las «hipótesis rivales aceptables» que han requerido el uso rutinario de grupos especiales de control actúan a modo de leyes empíricas bien establecidas: por ejemplo, los efectos de la práctica para el agregado de un grupo de control al diseño 2, la sugestibilidad para el falso grupo de control, el *shock* quirúrgico para el control con operación simulada, etc. Las hipótesis rivales son creíbles en la medida en que pueda atribuírseles categoría de leyes empíricas. Cuando en un cuasiexperimento se carece de controles, al interpretar los resultados hay que considerar bien la posibilidad de que tales resultados obedezcan a factores no tomados en cuenta. Cuanto más improbable sea esta posibilidad, más «válido» será el experimento.

Como lo señalamos al exponer el diseño de cuatro grupos de Solomon, cuanto más numerosas e independientes sean las formas en que se demuestra el efecto experimental, menos numerosas y probables se tornan todas las demás hipótesis rivales invalidantes. Se apela entonces a la economía. La «validez» del experimento viene a ser, pues, la de la admisibilidad relativa de las teorías rivales: la teoría de que X tuvo un efecto frente a las teorías de causación que comprenden los factores no controlados. Si cabe explicar la totalidad de varios conjuntos de diferencias por la hipótesis única de que X tiene un efecto, al paso que es necesario hipotetizar varios efectos separados

de variables no controladas, una para cada diferencia observada, entonces el efecto de X viene a ser el más defendible. Es frecuente recurrir a este modo de inferencia cuando los científicos tienen que limitarse a resumir literatura por carecer de experimentos perfectamente controlados. Así, Watson [1959, pág. 296] halló confirmatoria la evidencia de los efectos nocivos de la privación materna, porque se la ve confirmada por una amplia variedad de datos, cuyas insuficiencias específicas varían de unos estudios a otros. A su vez, Glickman [1961], a pesar de la presencia de hipótesis rivales sostenibles en cada uno de los estudios, consideró importantes las pruebas de un proceso de consolidación sólo porque la hipótesis rival sostenible variaba de un estudio a otro. Esta forma de inducción lógica, adoptada por lo común en la combinación de inferencias de distintos estudios, se introduce deliberadamente *dentro* de ciertos diseños cuasiexperimentales, en especial los «remendados», como el 15.

El recurrir a la economía no se justifica desde el punto de vista deductivo, sino que constituye más bien un supuesto general acerca de la naturaleza del mundo, que fundamenta casi toda aplicación de la teoría en la ciencia, por más que en aplicaciones particulares resulte a menudo errónea. En relación con esta observación hay otro argumento de admisibilidad, que invocaremos acaso más en detalle a propósito del muy utilizado diseño 10 (un buen diseño *cuasi*experimental, que a menudo se confunde con el verdadero diseño 4). Es la presunción de que, en casos de ignorancia, el efecto principal de una variable debe juzgarse más probable que la interacción de otras dos variables; o que, en general, los efectos principales son más probables que las interacciones. En su máxima expresión, cabe señalar que si cada interacción de orden superior es significativa y cada efecto es específico de determinados valores en todas las demás dimensiones posibles de tratamiento, ya no hay lugar para la ciencia. Si podemos generalizar alguna vez, es porque podemos hacer caso omiso de un gran cúmulo de factores potenciales determinantes.

Esto fue denominado por Underwood [1957*b*, pág. 6] «supuesto de causación finita». En otro lugar [1954], el mismo autor ha registrado la frecuencia de efectos principales y de interacciones en el *Journal of Experimental Psychology*, confirmando la relativa escasez de interacciones significativas (aunque las correcciones introducidas por el editor, tendientes a presentar resultados claros, nos hacen dudar de este hallazgo).

En los párrafos siguientes expondremos primero los experimentos con un solo grupo. Desde 1920, por lo menos, el diseño experimental predominante en psicología y educación ha sido el de grupo de control, como el 4, 6 o, acaso más a menudo aún, el diseño 10, que veremos más adelante. En las ciencias sociales, y considerando situaciones sobre el terreno, los diseños de grupo de control han predominado a tal punto que para algunos son sinónimo de experimentación. A consecuencia de ello muchos investigadores llegan a abandonar todo intento de experimentación en situaciones en las cuales no se disponga de grupos de control, terminando así como una imprecisión innecesaria. En realidad, varios diseños cuasiexperimentales aplicables a grupos aislados podrían emplearse provechosamente, y seguir los cánones lógicos e interpretativos experimentales, en muchos casos en que es imposible el diseño con grupo de control. La cooperación y la posibilidad de experimentar se dan a menudo en unidades administrativas naturales: una maestra dispone de su clase; el director de una escuela secundaria tal vez esté dispuesto a realizar encuestas periódicas sobre el estado de ánimo de los alumnos, etc. En tales situaciones el tratamiento diferencial de segmentos dentro de la unidad administrativa (requerido para el experimento con grupo de control) quizá resulte imposible en sentido administrativo o, aun cuando ello no ocurra, sea indeseable como experimento a causa de los efectos reactivos de los dispositivos. Para situaciones de esta índole bien podrían adoptarse experimentos con un grupo único.

7. Experimento de series cronológicas

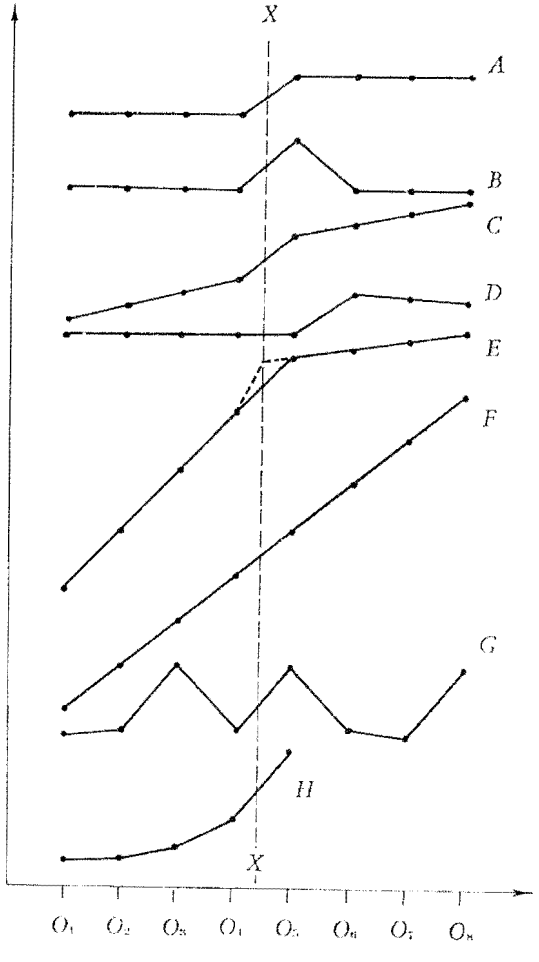
El diseño de series cronológicas consiste, en lo esencial, en un proceso periódico de medición sobre algún grupo o individuo y la introducción de una variación experimental en esa serie cronológica de mediciones, cuyos resultados se indican por medio de una discontinuidad en las mediciones registradas en la serie. Se lo puede diagramar de la manera siguiente:

$$O_1\ O_2\ O_3\ O_4 X O_5\ O_6\ O_7\ O_8$$

Este diseño experimental tipificó gran parte de la experimentación clásica del siglo XIX en las ciencias físicas y biológicas. Por ejemplo, si una barra de hierro cuyo peso no ha variado durante muchos meses se sumerge y retira de un baño de ácido nítrico, la deducción que relacionaría esta operación con la pérdida de peso de la barra seguiría alguna lógica experimental de esta índole. Por supuesto, pueden haber existido «grupos de control» de barras de hierro que se dejaron en los estantes y no perdieron peso, pero la medición e información de esos pesos constituiría un caso típico en el cual no se la consideraría ni necesaria ni pertinente. Parece, pues, probable que ese diseño experimental se considere a menudo válido en las ciencias de más éxito, si bien rara vez se lo acepte en las enumeraciones de diseños experimentales disponibles en las ciencias sociales. [Véase, sin embargo, Maxwell, 1958; Underwood, 1957*b*, pág. 133.] Hay buenas razones que justifican esa diferencia de categorías, y una cuidadosa consideración de ellas ofrecerá una mejor comprensión de las condiciones en que los científicos sociales podrían emplear con provecho el diseño cuando no hay modo de utilizar un control experimental más preciso. El diseño es típico de los experimentos clásicos del British Industrial Fatigue Research Board sobre factores que influyen en la producción industrial [p. ej., Farmer, Brooks y Chambers, 1923].

La figura 3 indica algunas posibles situaciones resultantes en series cronológicas en las cuales se había introducido una alteración experimental, según se indica por medio de la línea vertical *X*. Supongamos, a los fines de este estudio, que sentimos la tentación de deducir que *X* tuvo algún efecto en las series cronológicas con resultados como *A* y *B*, y quizá *C*, *D* y *E*, pero no un efecto en las series cronológicas tal como *F*, *G* y *H*, aunque el salto de valores de O_4 a O_5 fuese tan grande y desde el punto de vista estadístico tan persistente como, por ejemplo, las diferencias O_4 a O_5 en *A* y *B*. Aunque dejaremos el análisis del problema de las pruebas estadísticas para algunas páginas más adelante, se supone que el problema de la validez interna se reduce en definitiva a la cuestión de hipótesis competitivas aceptables que ofrezcan otras explicaciones probables, distintas del efecto de *X*, acerca del desplazamiento en las series cronológicas. Ofrecemos en el cuadro 2 un intento de lista de comprobación de los controles suministrados por este experimento en las mencionadas condiciones óptimas de resultado. Las ventajas del diseño de series cronológicas resultan muy evidentes en contraste con el diseño 2, con el que guarda una similitud superficial, ya que carece de grupo de control y utiliza mediciones previas y posteriores.

Figura 3. *Posibles configuraciones de los resultados de introducir una variable experimental en el punto X, en una serie cronológica de mediciones, $O_1 - O_8$. Salvo en el caso D, la diferencia $O_4 - O_5$ es la misma para todas las series cronológicas, en tanto que la legitimidad de inferir un efecto varía mucho, siendo máxima en A y B y totalmente injustificada en F, G y H.*



Observando en el cuadro 2 la lista de problemas de validez interna, vemos que la imposibilidad de controlar la historia es el más grave inconveniente del diseño 7. Es decir que existe la hipótesis rival de que no sea *X* sino otro acontecimiento más o menos simultáneo el que provocó el desplazamiento. Sobre la admisibilidad de eliminar tales estímulos externos debe basarse en todos los casos la confianza depositada en la interpretación de este experimento. Analicemos un experimento que exija mediciones reiteradas y el efecto que un filme documental produce sobre el optimismo o pesimismo de los alumnos con relación a la probabilidad de que estalle o no una guerra. En tal caso, no suministrar un control definido sobre la *historia* parecería sin duda muy grave, ya que es obvio que los estudiantes están todos los días expuestos a multitud de fuentes de estímulo en ese mismo sentido, además de las que maneja el experimentador en el aula. Por supuesto que, aun así, si el experimento se complementara con un cuidadoso registro de estímulos no experimentales de alguna relevancia, sería quizás aceptable una interpretación por la cual se justificara llevarlo a cabo. Como ya hemos dicho, la variable *historia* es la contrapartida de lo que en el laboratorio físico y biológico ha sido denominado *aislamiento experimental*. La admisibilidad de la *historia* como explicación de desplazamientos del tipo de los verificados en las series cronológicas *A* y *B* de la figura 3 depende, en gran parte, de la medida del aislamiento experimental que pueda conseguir el experimentador. Los estudios sobre reflejos condicionados realizados por Pavlov con perros, y que eran en definitiva experimentos de «un grupo» o «un animal», hubiesen sido mucho menos admisibles como fundamento de las teorías del sabio ruso si, en vez de haberlos efectuado en un laboratorio alejado de todo ruido, los hubiese practicado en cualquier esquina de la ciudad. Que es lo que constituye un aislamiento experimental varía con el problema que se estudia y el tipo de elementos de medición que se utilizan. Se exigen, sin duda, más precauciones para establecer el aislamiento experimental necesario cuando se han de estudiar partículas subatómicas en cámara de niebla o con un contador de centelleo, que para el experimento hipotético acerca del peso de las barras de hierro bañadas en ácido nítrico. En muchas situaciones que permitirían utilizar el diseño 7, sería admisible que el experimentador sostuviese haber trabajado en condiciones de aislamiento experimental, si tuvo conciencia de los posibles acontecimientos rivales también capaces de producir el cam-

Cuadro 2. *Fuentes de invalidación para los diseños 7 a 12.*

	Fuentes de invalidación											
	Interna								Externa			
	Historia	Maduración	Administración de tests	Instrumentación	Regresión	Selección	Mortalidad	Interacción de selección y maduración, etc.	Interacción de administración de tests y X	Interacción de selección y X	Dispositivos reactivos	Interferencia de X múltiples
Diseños cuasiexperimentales:												
7. Series cronológicas $O\ O\ O\ OXO\ O\ O\ O$	−	+	+	?	+	+	+	+	−	?	?	
8. Diseño de muestras cronológicas equivalentes $X_1O\ X_2O\ X_3O$ X_4O, etc.	+	+	+	+	+	+	+	+	−	?	−	−
9. Diseño de muestras materiales equivalentes $M_aX_1O\ M_bX_2O$ M_cX_3O M_dX_4O, etc.	+	+	+	+	+	+	+	+	−	?	?	−
10. Diseño de grupo de control no equivalente $O\ X\ O$ $O\ \ O$	+	+	+	+	?	+	+	−	−	?	?	
11. Diseños compensados $X_1O\ X_2O\ X_3O\ X_4O$ $X_2O\ X_3O\ X_1O\ X_3O$ $X_3O\ X_1O\ X_4O\ X_2O$ $X_4O\ X_2O\ X_3O\ X_1O$	+	+	+	+	+	+	+	?	?	?	?	−
12. Diseño de muestra separada pretest-postest $R\ O\ (X)$ $R\ \ X\ O$	−	−	+	?	+	+	−	−	+	+	+	
12*a* $R\ O\ (X)$ $R\ \ X\ O$ $R\ \ O\ (X)$ $R\ \ X\ O$	+	−	+	?	+	+	−	+	+	+	+	
12*b* $R\ O_1\ \ (X)$ $R\ \ O_2\ (X)$ $R\ \ X\ O_3$	−	+	+	?	+	+	−	?	+	+	+	
12*c* $R\ O_1\ X\ O_2$ $R\ \ X\ O_3$	−	−	+	?	+	+	+	−	+	+	+	

bio, y pudo descartar con suficiente lógica la probabilidad de que estos últimos lo explicaran.

Entre otras variables externas que, por razones de conveniencia, podrían incluirse en el factor *historia*, están los efectos meteorológicos. Los experimentos de esa índole (p. ej., los estudios sobre rendimiento laboral) tal vez se prolonguen durante lapsos que abarquen cambios estacionales, y entonces las fluctuaciones estacionales en iluminación, condiciones meteorológicas, etc., pueden confundirse con la introducción de variaciones experimentales. Quizá fuera mejor incluir en la *historia*, aunque sean en cierto sentido análogos a la *maduración*, los desplazamientos periódicos de las series cronológicas referidas a las costumbres institucionales del grupo, como los ciclos semanales de trabajo y de pagos de salarios, los períodos de exámenes y vacaciones, y las fiestas escolares. Las series de observaciones deberían ordenarse de tal manera que se mantuvieran constantes los ciclos conocidos, o de lo contrario ser lo bastante prolongadas como para incluir en su totalidad varios de esos ciclos.

Continuemos con los factores que hay que controlar: la *maduración* parece quedar eliminada sobre la base de que, si el resultado es como los de los ejemplos *A* y *B* de la figura 3, ella no ofrece de ordinario hipótesis rivales aceptables para explicar algún desplazamiento producido entre O_4 y O_5 que no se había dado en los anteriores períodos observados. (No obstante, la maduración no siempre es uniforme y regular. Nótese cómo la súbita iniciación de las menstruaciones en alumnas del primer año de la escuela secundaria podría aparecer en un diseño 7 como un efecto del cambio de escuelas sobre los registros fisiológicos, si no supiésemos que no era así.) Asimismo, la aplicación de tests parece, en general, hipótesis rival no creíble para un salto entre O_4 y O_5. Si solo tuviéramos las observaciones en O_4 y O_5, como en el diseño 2, careceríamos de ese medio de convertir en inaceptables los efectos de maduración y los tests-retests. Esta es la gran ventaja del diseño 7 sobre el diseño 2.

De igual modo, muchas hipótesis que invocarían variaciones en la *instrumentación* carecerían de base racional específica para suponer que en aquella ocasión particular, a diferencia de otras anteriores, se había producido el error de los aparatos de medición. No obstante, el signo de interrogación en el cuadro 2 llama la atención recordando las posibles situaciones en que un cambio en la calibración del instrumento de medición podría dar lugar a que se lo interpretase como efec-

to de X. Si el procedimiento de medición implica las apreciaciones de observadores humanos conocedores del plan experimental, puede producirse una seudoconfirmación de la hipótesis a causa de las expectativas del observador. Así, el cambio experimental de poner en posesión de su cargo a un nuevo director puede producir una variación en las estadísticas de faltas disciplinarias, en vez de influir sobre la tasa de infracciones como tal. A menudo puede emplearse el diseño 7 para medir los efectos de un importante cambio introducido en la política administrativa. Teniendo esto en cuenta, convendría evitar el cambio de instrumentos de medición a la vez que se modifica la política. En la mayor parte de los casos sería preferible, a fin de preservar la interpretabilidad de una serie cronológica, continuar empleando dispositivos un tanto anticuados, en vez de sustituirlos por otros más modernos pero distintos.

Los efectos de la *regresión* suelen consistir en una función negativamente acelerada del tiempo trascurrido, razón por la cual no son aceptables como explicaciones de un efecto en O_5 mayor que los efectos en O_2, O_3 y O_4. La *selección* como fuente de efectos principales se elimina tanto en este diseño como en el 2, si en todas las O están implicadas las mismas personas. Si en un determinado grupo los datos se recopilan sobre la base de sus integrantes individuales, se puede eliminar la mortalidad en ese experimento lo mismo que en el diseño 2. No obstante, si las observaciones se refieren a datos colectivos, habría que llevar un registro del ausentismo, las renuncias y las reposiciones, a fin de asegurarse de que las coincidencias de cambios de personal no ofrezcan hipótesis rivales aceptables.

En cuanto a la validez externa, está claro que el efecto experimental bien podría ser específico para las poblaciones sujetas a reiteración de pruebas. Es improbable que ello constituya una limitación a la investigación sobre la enseñanza en establecimientos escolares, a menos que el experimento se realice con O artificiales no comunes en la situación escolar habitual. Además, este diseño es de particular aplicación en las situaciones institucionales en que se llevan registros regulares que constituyen, por tanto, parte natural del ambiente en que se realiza el experimento. Las pruebas anuales de calificación en las escuelas públicas, los registros de enfermedad, etc., no son por lo común reactivos, puesto que son típicos del universo al cual se quiere hacer la generalización. La interacción *selección-X* se refiere a la restricción de los efectos

de la variable experimental a esa muestra específica, así como a la posibilidad de que esa reacción no fuese típica de algún universo de interés más general, con respecto al cual el grupo expuesto, constituido naturalmente, sea una muestra sesgada. Por ejemplo, la necesidad de datos puede limitarnos a aquellos alumnos que han tenido asistencia perfecta durante largos períodos: un evidente subconjunto selecto. Además, si se han utilizado O nuevas, ese reiterado acontecimiento puede haber provocado ausentismo.

Para que estas series cronológicas se interpreten como experimentos, es imprescindible que el experimentador especifique de antemano la relación cronológica esperada entre la introducción de la variable experimental y la manifestación de un efecto. Si se lo hubiese hecho así, la situación indicada en la serie cronológica D de la figura 3 podría ser tan concluyente como la de A. Las encuestas exploratorias que decidiesen según las circunstancias sobre las interpretaciones de efecto diferido, requerirían una validación cruzada antes de que fueran interpretables. A medida que aumenta el lapso trascurrido entre X y el efecto, aumenta asimismo la admisibilidad de los efectos resultantes de acontecimientos históricos externos. También parece imprescindible que se especifique la X antes de examinar el resultado de la serie cronológica. El examen *post hoc* de una serie cronológica para inferir cuál fue la X que precedió al desplazamiento más notable tiene que descartarse, a causa de que el aprovechamiento oportunista del azar que él permite dificulta, y hasta imposibilita, cualquier intento de comprobar la significación de los efectos.

La preponderancia de este diseño en las ciencias más prósperas debería hacernos sentir algún respeto por él; recuérdese, sin embargo, que los aspectos de «aislamiento experimental» y las «condiciones constantes» lo hacen más interpretable para ellos que para nosotros. Recuérdese, asimismo, que en el uso que suele hacerse de él, un experimento único jamás es concluyente. Aunque puede ocurrir que no se utilice un grupo de control, antes de establecer un principio es menester que varios investigadores repitan el diseño 7 en muchos lugares distintos. Tal debería ser, asimismo, la aplicación que nosotros le diéramos a este diseño. Deberíamos utilizarlo allí *donde no se pueda hacer nada mejor controlado.* Organizaremos nuestra contabilidad institucional de tal modo que nos suministre el mayor número posible de series cronológicas para dichas evaluaciones, y trataremos de examinar con mayor cuidado que hasta entonces los efectos de los cambios admi-

nistrativos y otros acontecimientos súbitos y arbitrarios como X. Pero no los consideraremos definitivos hasta haberlos repetido una y otra vez en situaciones diversas.

Tests de significación para el diseño de serie cronológica

Si las ciencias más avanzadas no emplean tantos tests de significación como la psicología y la pedagogía es, sin duda, porque la magnitud y claridad de los efectos con que trabajan son tales que los hacen innecesarios. Si se aplicase en ellas nuestros tests habituales de significación, se descubrirían también allí elevados índices de este factor. Parece, sin embargo, típico de la ecología de las ciencias sociales tener que trabajar con minerales pobres, para los que no son necesarias las pruebas de significación. También es probable que allí donde el sentido común o las apreciaciones intuitivas señalen con claridad un efecto resulte posible, por lo común, alguna prueba de significación que ratifique las consideraciones en que se funda el juicio intuitivo. Así, se pueden aplicar pruebas de significación sobre los efectos de X que distinguirían entre los varios resultados que ilustra la figura 3, juzgando que A y B son significativos y que F y G no lo son. Veamos algunos posibles enfoques.

Ante todo debemos rechazar, por inadecuados, ciertos tratamientos imaginables. Si la información recogida en la figura 3 representa medias grupales, será insuficiente una simple prueba de significación de la diferencia entre las observaciones de O_4 y O_5. Aunque en las series F y G estas suministrasen relaciones t de elevada significación, no podríamos decir que los datos demostraban el efecto de X a causa de la presencia de otros desplazamientos significativos similares, que a veces ocurren, y para los cuales no tendríamos explicación experimental alguna que les correspondiese. Cuando se trabaja con la información obtenida de encuestas nacionales de opinión, suelen encontrarse desplazamientos muy significativos entre una y otra consulta que, desde el punto de vista del científico encargado de las interpretaciones, no son más que «ruidos» aleatorios, ya que constituyen una parte de la variación en los fenómenos para la que se carece de explicación. A fin de que sea interpretable, el efecto de un bien perfilado acontecimiento o variable experimental debe trascender ese nivel ordinario de desplazamiento. Asimismo, una prueba de significación que abarque los datos combinados de todas las observaciones pre-X y post-X resulta inadecuada, pues no distinguiría entre casos del tipo F y casos del tipo A.

En el desarrollo de un test de significación suele haber una enojosa falta de independencia. Si tal carencia estuviese distribuida en forma homogénea entre todas las observaciones, ella dejaría de constituir una amenaza para la validez interna, aunque fuera una limitación a la validez externa. Lo que sí resulta inoportuno es que en casi todas las series cronológicas se encontrará que las observaciones adyacentes son más similares que las no adyacentes (es decir, que la autocorrelación del defasaje 1 es mayor que la del 2, etc.). Así, una influencia o perturbación aleatoria externa que influya sobre un punto de observación, por ejemplo, en O_5 u O_6, también alterará O_7 y O_8, por lo cual no se las puede tratar como varias desviaciones independientes de la extrapolación de la tendencia $O_1 - O_4$.

El test de significación utilizado dependerá en parte de la índole hipotética del efecto de X. Si está implicado un modelo como la línea B, se podría utilizar una prueba de la desviación de O_5 respecto de la extrapolación de $O_1 - O_4$. Mood [1950, págs. 297-98] ofrece una prueba de esta índole. Ese test podría emplearse en todos los casos, pero parecería innecesariamente insuficiente si se supone una mejoría constante o un mayor índice de ganancia. Para esos casos, sería aconsejable una prueba que emplease todos los puntos. Son dos los componentes que constituirían tales pruebas de significación: la intersección y la pendiente. Por intersección queremos significar el salto en la serie cronológica en el punto X (o en algún desplazamiento especificado después de X). Así las líneas A y C indican un desplazamiento de intersección sin variación en la pendiente. La línea E ilustra un cambio en la pendiente pero no en la intersección, ya que coinciden las extrapolaciones pre-X a X y post-X a X. A menudo ambas se cortan, y la pendiente quedaría modificada por una X efectiva. Podría conseguirse una prueba pura de intersección en forma análoga a la aplicación de una prueba Mood desde una y otra dirección a la vez. En este caso estarían implicados dos puntos extrapolados, con observaciones pre y post X extrapoladas a un punto X intermedio entre O_4 y O_5. Las pruebas estadísticas casi con seguridad comprenderían, en todas las series cronológicas (con excepción de las más extensas) ajustes lineales a los datos, tanto por conveniencia como porque un ajuste más exacto agotaría los grados de libertad, no dejando oportunidad alguna para verificar la hipó-

tesis de cambio. Sin embargo, en muchos casos la presunción de linealidad puede no ser correcta. La admisibilidad de inferir un efecto de X es mayor en un punto próximo a X. Cuanto más gradual o más lento sea el efecto supuesto, más grave será la confusión con la historia, ya que aumenta el número de posibles causas externas.

8. Diseño de muestras cronológicas equivalentes

La forma más común de diseño experimental utiliza una muestra equivalente de personas a fin de suministrar la base con la cual comparar los efectos de la variable experimental. Como contraste, una forma recurrente de experimentación con un solo grupo utiliza dos muestras equivalentes de sesiones, con la variable experimental en una de ellas y no en la otra. Ese diseño puede esquematizarse de la siguiente manera (aunque la intención es obtener una alternación aleatoria, no regular):

$$X_1O \quad X_0O \quad X_1O \quad X_0O$$

Este diseño puede considerarse una forma del experimento de serie cronológica con la introducción reiterada de la variable experimental. El experimento es, sin duda, de mayor utilidad cuando se anticipa que el efecto de la variable experimental será de carácter transitorio o reversible. Al paso que la lógica del experimento puede considerarse una extensión del experimento de serie cronológica, el modo de análisis estadístico tiene mayor similitud, en general, con el del experimento de dos grupos en los cuales se emplea la significación de la diferencia entre las medias de dos conjuntos de mediciones. Por lo común, las mediciones están apareadas de manera muy específica con las presentaciones de la variable experimental, siendo a menudo concomitante, como en los estudios de aprendizaje, rendimiento laboral, condicionamiento, reacción fisiológica, etc. Acaso el más típico de los primeros usos de este diseño experimental, como en los estudios de Allport [1920] y Sorokin [1930] sobre el rendimiento escolar en condiciones diversas, consistió en la comparación mutua de dos variables experimentales, es decir X_1 *versus* X_2, en vez de una sola de ellas con un control. Para la mayoría de los fines resultan ineficaces la simple alternación de condiciones y el empleo de

un espaciamiento temporal constante, sobre todo cuando pueden introducir un elemento de confusión con un ciclo diario, semanal o mensual, o cuando a causa de la periodicidad predecible, un condicionamiento indeseable al intervalo temporal puede intensificar la diferencia entre las distintas presentaciones. Así, Sorokin se aseguró de que cada tratamiento experimental se realizara con idéntica frecuencia de mañana y de tarde.

Casi todos los experimentos con este diseño han empleado en proporción pocas repeticiones de cada condición experimental, pero una extensión de la teoría del muestreo como la representada por Brunswik [1956] señala la necesidad de grandes muestreos aleatorios, representativos y equivalentes, de los períodos. Kerr [1945] es quien más se ha aproximado tal vez a este ideal en sus experimentos acerca del influjo de la música sobre el rendimiento industrial. Cada uno de ellos comprendió un solo grupo experimental con una muestra aleatorizada y equivalente de días a lo largo de varios meses. De esta forma, en un experimento pudo comparar 56 días con música y 51 días sin ella, y en otro tres tipos diferentes de música, representada cada una por muestras equivalentes de 14 días.

Tal como lo empleó Kerr, por ejemplo, el diseño 8 parece en general internamente válido. La *historia*, que es el principal inconveniente del experimento con series cronológicas, se controla presentando X en numerosas sesiones separadas, haciendo así improbable en extremo cualquier otra explicación fundada en la coincidencia de acontecimientos externos. Las otras fuentes de invalidación se controlan con la misma lógica detallada a propósito del diseño 7. En cuanto a la validez externa, es evidente que solo cabe extender la generalización a poblaciones probadas con frecuencia. El efecto reactivo de los dispositivos y la conciencia de que se es objeto de la experimentación constituyen una deficiencia de esta prueba. Cuando son grupos separados los que reciben las distintas X, puede ocurrir (sobre todo en el diseño 6) que ignoren por completo la existencia del experimento o de los tratamientos que se comparan. No ocurre así cuando se maneja un solo grupo y se lo expone en repetidas sesiones a una u otra condición, por ejemplo, a una base de cómputo de pago contra otra en el experimento de Sorokin; una condición de trabajo contra otra en el de Allport; un tipo de ventilación contra otro en los estudios de Wyatt, Fraser y Stock [1926], y una clase de música contra otra en el de Kerr (aunque este

investigador tomó cuidadosas precauciones para conseguir que una programación variada se convirtiese en parte integrante del ambiente laboral). En cuanto a la interacción de *selección* y X, se da, como es habitual, la limitación de la generalización de los efectos demostrados de X al tipo particular de la población de que se trata.

Este diseño experimental lleva implícito un riesgo para la validez externa que se encontrará en todos los experimentos descritos en este trabajo en los cuales se presentan muchos niveles de X para el *mismo* conjunto de personas. Ese efecto se ha denominado «interferencia de X múltiples». El efecto de X_1, en la situación más simple, en que se la compara con X_0, sólo puede generalizarse a condiciones de presentaciones repetidas y espaciadas de X_1. No se ofrece una base sólida para la generalización a posibles situaciones en que X_1 esté siempre presente, o a la condición en que se la introduzca en una sola sesión. Además, la condición X_0 o la ausencia de X no es típica de períodos sin X en general, sino que es representativa solo de ausencias de X intercaladas entre presencias de este factor. Si X_1 tiene algún efecto prolongado que llega a influir en los períodos sin X, como parece por lo común probable, el diseño experimental, comparado con un estudio con diseño 6, por ejemplo, puede subestimar el efecto de X_1. Por el contrario, el hecho mismo de que se produzcan frecuentes desplazamientos puede incrementar el valor de estímulo de una X, excediendo al que se daría en una presentación continua y homogénea. En el estudio de Kerr las melodías hawaianas influirían sobre el trabajo de manera bastante diferente si se las intercalase durante todo un día entre otras formas de música, que si constituyen el único «alimento» musical. Los diseños experimentales de Ebbinghaus [1885] pueden considerarse en lo esencial de esta índole y, como lo ha destacado Underwood [1957*a*], las leyes por él descubiertas están limitadas en sus posibilidades de generalización a una población de personas que hayan aprendido docenas de otras listas muy similares. Incluso gran parte de sus descubrimientos no se verifican en personas que aprenden una sola lista de sílabas desprovistas de significado. Así, mientras el diseño es internamente válido, su validez externa suele verse limitada en gran parte por ciertos tipos de contenido. [Véase también Kempthorne, 1952, cap. 29.]

Nótese, sin embargo, que muchos aspectos de la enseñanza sobre los cuales se desearía experimentar pueden muy bien tener efectos restringidos, para los fines prácticos, al período

de presencia concreta de X. Para esos objetivos, este diseño podría ser muy valioso. Supongamos que un maestro pone en tela de juicio el valor de las lecciones en voz alta contra el del estudio individual en silencio. Variando esos dos procedimientos durante una serie de unidades de lecciones, se podría preparar un experimento interpretable. De ese modo cabría estudiar el efecto de la presencia en el aula de un padre que actuara como observador durante un debate voluntario entre los alumnos. El conocimiento de ese tipo de diseños puede poner al alcance de un maestro individual la verificación experimental de las alternativas. Esto podría dar lugar a procedimientos de tipo piloto que, de resultar promisorios, se examinarían por medio de experimentos de mayor envergadura y mejor coordinados.

Este enfoque es aplicable a un muestreo de sesiones con un solo sujeto. Aunque no es habitual todavía administrar tests de significación, es este un diseño muy utilizado en la investigación fisiológica, en la cual se aplica repetidas veces un estímulo a un animal, poniendo sumo cuidado en evitar cualquier periodicidad en la estimulación, ya que este último aspecto corresponde al requisito de aleatorización para aquellas sesiones en que así lo demande la lógica del diseño. También pueden utilizarse cuadrados latinos en vez de la aleatorización simple [p. ej., Cox, 1951; Maxwell, 1958].

Tests de significación para el diseño 8

Una vez más necesitamos pruebas de significación apropiadas para este tipo particular de diseño. Adviértase que hay implícitas en él dos dimensiones de generalización: con respecto a las sesiones y con respecto a las personas. Si consideramos un caso en que se utilice una sola persona, es obvio que la generalización de la prueba de significación se limitará a esa persona en particular, comprendiendo una generalización entre casos, para cuyo fin convendrá utilizar una t con un número de grados de libertad igual al de sesiones menos dos. Si se poseen registros individuales de cierto número de personas sometidas al mismo tratamiento y todas ellas comparten el mismo grupo, se tendrán también datos para generalizar entre personas. En esta situación habitual dos estrategias parecen comunes. Una, errónea, es la de generar a propósito de cada individuo un puntaje único para cada tratamiento experimental, y aplicar luego tests de significación de la diferen-

cia entre las medias con datos correlacionados. Esta es la lógica de los análisis de Allport y Sorokin, aunque en realidad no se utilizaron tests de significación. Pero cuando solo están implicadas una o dos repeticiones de cada condición experimental, los errores de muestreo de las sesiones pueden ser muy grandes o el control de la historia muy deficiente. Los errores aleatorios en el muestreo de sesiones podrían constituir lo que a la luz de este análisis parecen ser diferencias significativas entre unos y otros tratamientos. Esto será un error muy grave si el efecto de las sesiones es significativo y apreciable. Sobre ese supuesto lógico se podría obtener, por ejemplo, una diferencia sumamente significativa entre X_1 y X_2, cuando cada una solo haya sido presentada una vez y cuando en una sesión algún acontecimiento externo haya producido por azar un resultado notable. Parece, pues, imprescindible que para cada tratamiento se «incluyan» por lo menos dos sesiones y estén representados los grados de libertad entre ellas. La mejor forma de cumplir con este requisito es, quizá, probar ante todo la diferencia entre las medias de tratamiento y un término de error entre las diversas sesiones y con respecto a cada tratamiento. Después de establecer así la significación del efecto del tratamiento, se podría proceder a determinar la proporción de sujetos para los cuales se verifica, obteniendo así datos sobre la posibilidad de generalizar el efecto a diversas personas. Las mediciones y muestreos repetidos de sesiones plantean muchos problemas estadísticos, algunos de los cuales no han sido resueltos todavía [Collier, 1960; Cox, 1951; Kempthorne, 1952].

9. Diseño de materiales equivalentes

El diseño 9 está íntimamente relacionado con el de muestras cronológicas equivalentes, y su argumento se funda en la equivalencia de las muestras de materiales a que se aplican las variables experimentales que se comparan. Siempre, o casi siempre, hay también implicadas muestras cronológicas equivalentes, pero pueden estar intercaladas en forma tan sutil o intrincada, que prácticamente vienen a constituir una equivalencia temporal. En un diseño con un grupo y X repetida, se requieren materiales equivalentes allí donde la índole de las variables experimentales sea tal que los efectos son permanentes, y los distintos tratamientos y repeticiones de ellos

deben aplicarse a un contenido no idéntico. El diseño puede expresarse así:

$$M_aX_1O \quad M_bX_0O \quad M_cX_1O \quad M_dX_0O \quad \text{etc.}$$

Las M indican materiales específicos, siendo la muestra M_a, M_c, etc., en términos de muestreo, igual a la muestra M_b, M_d, etc. La importancia de la equivalencia de muestreo de ambos conjuntos de materiales quedaría acaso mejor indicada si se diagramara el diseño de esta manera:

Una persona o grupo	{	Muestra de materiales $A(O)\ X_0\ O$
		Muestra de materiales $B(O)\ X_1\ O$

Las O entre paréntesis indican que en algunos diseños se utilizará un pretest y en otros no.

El experimento de Jost [1897] sobre práctica masiva contra práctica distribuida ofrece un magnífico ejemplo. En su tercer experimento se prepararon bastante al azar doce listas de doce sílabas carentes de sentido. Seis se asignaron a la práctica distribuida y seis a la masiva. Las doce se aprendieron simultáneamente en un lapso de siete días, combinándose con cuidado su programación de modo que se controlasen la fatiga y otros aspectos. Siete de aquellos conjuntos de seis listas distribuidas y seis masivas se aprendieron durante un lapso que se extendió desde el 6 de noviembre de 1895 hasta el 7 de abril de 1896. Al final, Jost obtuvo resultados sobre 40 listas diferentes de sílabas aprendidas con práctica masiva y 40 con práctica distribuida. La interpretabilidad de las diferencias descubiertas en el único sujeto de la prueba, G. E. Müller, depende de la equivalencia de muestreo de las listas no idénticas existentes. Dentro de estos márgenes, el experimento descripto parece tener validez interna. Los descubrimientos, naturalmente, se limitan a los rasgos psicológicos de Müller en 1895 y 1896 y al universo de material de memorización muestreado. Para poder generalizar a otras personas y establecer una ley psicológica más general, habría, por supuesto, que repetir el experimento con muchos individuos.

Otro ejemplo proviene de los primeros estudios sobre conformidad a la opinión del grupo. Moore [1921], por ejemplo, obtuvo una estimación «control» de estabilidad en retest de las respuestas a un conjunto de ítems de un cuestionario, después de lo cual comparó esa medida con la variación resultante cuando, con otro conjunto, se acompañó el retest con

una manifestación de la opinión de la mayoría. Consideremos en cambio un estudio en el cual se solicita de los alumnos que manifiesten su parecer acerca de un determinado número de temas presentados en un extenso cuestionario. Se dividen entonces las preguntas en dos grupos tan equivalentes como sea posible. En un momento posterior, se devuelven los cuestionarios a los alumnos y el grupo vota por cada uno de los ítems indicados. Se falsifican esos votos a fin de indicar mayorías opuestas a las que prevalecieron en las dos muestras de ítems. Como medición post-*X*, se solicita de los alumnos que vuelvan a votar sobre todos los temas. En caso de que el argumento de equivalencia de muestreo de ambos conjuntos de elementos fuera correcto, las diferencias de desplazamientos entre los dos tratamientos parecerían suministrar una prueba definitiva acerca de los efectos de dar a conocer las opiniones del grupo, aun en ausencia de grupo de control alguno.

A semejanza del diseño 8, el 9 tiene validez interna en todos los puntos, y en general por los mismos motivos. Obsérvese, a propósito de la validez externa, que en el diseño 9, como en todos los experimentos con mediciones repetidas, los efectos pueden ser bastante específicos de las personas medidas en varias sesiones. En pruebas de aprendizaje, las mediciones son parte tan integrante de la situación experimental propia del método típico utilizado en la actualidad (aunque no necesariamente en el método de Jost, en el cual las prácticas comprendieron cantidades controladas de lecturas de las listas), que esta limitación a la generalización pierde toda importancia. Parecería que en el diseño 9 hay menos posibilidades de dispositivos reactivos que en el 8 a causa de la heterogeneidad de los materiales y la mayor probabilidad de que los sujetos no adviertan que reciben tratamientos diferentes en momentos diferentes y para ítems diferentes. Esta escasa reactividad no aparecería en el experimento de Jost, pero sí en el estudio de conformidad. Es probable, pues, que la interferencia entre los niveles de la variable experimental o entre los materiales sea una innegable imperfección de este experimento, al igual que en el diseño 8.

Tenemos un ejemplo específico del tipo de limitación así introducido acerca de los descubrimientos de Jost. Este investigador informó que el aprendizaje espaciado era más eficaz que la práctica masiva. De las condiciones generales de su experimentación cabe inferir que estaba justificado al generalizar sólo para las personas que estuviesen aprendiendo muchas

listas, o sea, las que tenían un elevado nivel de interferencia. La investigación contemporánea indica que la superioridad del aprendizaje espaciado sólo se restringe a tales poblaciones, y que en personas que aprenden por primera vez materiales muy nuevos, no se da esa ventaja [Underwood y Richardson, 1958].

Estadísticas del diseño 9

Es obvio que el muestreo de materiales guarda relación con la validez y el grado de prueba del experimento. Como tal, es probable que la *N* para el cálculo de la significación de las diferencias entre las medias de grupos de tratamiento debiera haber sido una *N* de listas en el experimento de Jost (o una *N* de elementos en el estudio de conformidad) a fin de que se representara ese importante campo de muestreo. Se lo debe completar con una base de generalización entre personas. En la actualidad, acaso lo mejor sea hacerlo en forma seriada, estableciendo ante todo la generalización entre la muestra de listas o ítems, computando después un puntaje de efectos experimentales para cada persona, y empleando todo ello como base para la generalización entre personas. (Véase la bibliografía antes citada, a propósito del diseño 8, con respecto a las precauciones que deben tomarse.)

10. Diseño de grupo de control no equivalente

Uno de los diseños experimentales más difundidos en la investigación educacional comprende un grupo experimental y otro de control, de los cuales ambos han recibido un pretest y un postest, pero no poseen equivalencia preexperimental de muestreo. Por lo contrario, los grupos constituyen entidades formadas naturalmente (como una clase, por ejemplo) tan similares como la disponibilidad lo permita, aunque no tanto, sin embargo, que se pueda prescindir del pretest. La asignación de *X* a uno u otro grupo se supone aleatoria y controlada por el experimentador.

$$\begin{array}{ccc} O & X & O \\ O & & O \end{array}$$

Dos cosas han de tenerse claras sobre este diseño. Ante todo, que no se lo debe confundir con el 4, el diseño con grupo de control pretest-postest, donde los sujetos experimentales que se toman de una población común se asignan *en forma aleatoria* al grupo experimental y de control. En segundo lugar, que, a pesar de ello, hay que admitir que el diseño 10 es utilizable en muchas oportunidades en que son imposibles los diseños 4, 5 o 6. Sobre todo, habrá que reconocer que aun el agregado de un grupo de control no equiparado o no equivalente reduce en gran parte la ambigüedad de las interpretaciones que derivan del diseño 2 de un grupo pretest-postest. Cuanto más similares sean en su reclutamiento el grupo experimental y el de control y más se confirme esa similitud por los puntajes del pretest, más eficaz resulta ese control. Suponiendo que estos ideales se aproximen a los objetivos de la validez interna, podemos considerar que el diseño controla los principales efectos de la historia, la maduración, la administración de tests y la instrumentación, donde la diferencia para el grupo experimental entre el pretest y el postest (si fuera mayor que para el grupo de control) no puede explicarse por efectos principales de esas variables, como los que afectarían tanto al grupo experimental como al de control. (Sin embargo, deben extremarse las precauciones sobre la historia intrasesional mencionadas en el diseño 4.)

Un esfuerzo por explicar una ganancia pretest-postest propia del grupo experimental en términos de factores externos, como historia, maduración o aplicación de tests, tiene que suponer una interacción entre esas variables y las diferencias específicas de selección que se den entre el grupo experimental y el de control. Aunque tales interacciones son en general poco probables, hay un cierto número de situaciones en las que podrían invocarse. Acaso las más comunes sean las interacciones que implican *maduración*. Si el grupo experimental consta de pacientes de psicoterapia y el de control de alguna otra población disponible a la cual se le hayan administrado un test y un retest, una ganancia peculiar al grupo experimental bien podría interpretarse como un proceso espontáneo de remisión típico de grupo tan extremo, ganancia que se hubiese producido también aun en ausencia de *X*. Tal interacción entre selección y maduración (o selección-historia, o selección-test) podría confundirse con el efecto de *X*, constituyendo por tanto una amenaza a la validez *interna* del experimento. Esta posibilidad ha sido representada en la octava

columna del cuadro 2 y es el principal factor de validez *interna* que caracteriza a los diseños 4 y 10.

Acaso se aclare este punto con un ejemplo concreto de investigación educacional. El estudio de Sanford y Hemphill [1952] sobre los efectos de un curso de psicología en Annapolis ofrece una excelente ilustración del diseño 10. En ese trabajo, el Segundo Curso de Annapolis constituyó el grupo experimental, y el Tercero, el de control. Las mayores ganancias registradas por el grupo experimental podrían explicarse como parte de un proceso general de perfeccionamiento, con resultados máximos en los primeros dos cursos y mínimos en el tercero y cuarto, constituyendo, por tanto, una interacción entre los factores de selección que diferencian los grupos experimental y de control y las variaciones naturales maduración características de tales grupos, y no un efecto del programa experimental. El grupo particular de control utilizado por Sanford y Hemphill posibilita alguna verificación de esta interpretación rival (en forma un tanto similar al diseño 15, que expondremos más adelante). La hipótesis de selección-maduración pronosticaría que el Tercer Curso (grupo de control) habría de indicar en su test inicial una superioridad respecto de las mediciones pretest del Segundo Curso (grupo experimental), con magnitud casi igual a la hallada entre el pretest y el postest de este último grupo. Por fortuna para la interpretación de su experimento, no ocurrió en general así. Las diferencias entre los cursos en el pretest no presentaban en la mayoría de los casos el mismo sentido ni igual magnitud que las ganancias pretest-postest del grupo experimental. Sin embargo, sus comprobaciones de una ganancia significativa para el grupo experimental en puntajes de confianza en el cuestionario de situaciones sociales pueden explicarse como un mecanismo artificial de selección-maduración. El grupo experimental pasó de 43,26 puntos a 51,42, en tanto que el Tercer Curso comenzó por un puntaje de 55,82 y continuó aumentando hasta alcanzar 56,78. La hipótesis de interacción entre selección y maduración será en ocasiones aceptable, aun cuando los grupos obtengan puntajes pretest idénticos. El más común de tales casos será aquel en que un grupo obtenga una tasa de maduración o variación autónoma más elevada que el otro. El diseño 14 ofrece una extensión del 10 que tendería a eliminar este factor.

El otro gran problema de la validez interna en el diseño 10 es la regresión. Como se indicó con «?» en el cuadro 2, cabe evitar ese riesgo, pero no siempre al tropezar con él se lo

sortea. En general, si se ha elegido cualquiera de los grupos de comparación por sus puntajes extremos de *O* o mediciones correlativas, una diferencia en el grado de desplazamiento de pretest a postest entre ambos grupos bien puede ser producto de la regresión y no efecto de *X*. Esta posibilidad ha tenido mayor trascendencia a causa de una obcecada y engañosa tradición en el ámbito de la experimentación educacional, por la que se considera la equiparación como una técnica apropiada y suficiente para establecer la equivalencia preexperimental de grupos. Este error ha ido acompañado por la falta de distinción entre los diseños 4 y 10 y los diferentes papeles representados por la equiparación en los puntajes de pretest en ambas condiciones. En el diseño 4, puede considerarse este procedimiento como un complemento provechoso de la aleatorización, pero no como un sustituto de ella; en términos de puntajes en el pretest o en las variables relativas, cabe organizar la población total disponible para fines experimentales en pares de sujetos cuidadosamente equiparados; los integrantes de esos pares se asignarán *al azar* a las condiciones experimentales o de control. Esa equiparación más la ulterior aleatorización suelen producir un diseño experimental más preciso que la aleatorización por sí sola.

No debe confundirse con ese ideal la técnica, correspondiente al diseño 10, de tratar de compensar las diferencias entre los grupos experimentales y de control no equivalentes mediante un procedimiento de equiparación, cuando no se puede hacer la asignación aleatoria a tratamientos. Si en el diseño 10 las medias de los grupos son sustancialmente diferentes, el proceso de equiparación, no solo no suministra la igualación pretendida, sino que provoca la presencia de efectos indeseados de regresión. Se torna previsible que ambos grupos diferirán en sus puntajes postest en forma por completo independiente de cualesquiera efectos de *X*, así como que esa diferencia variará en proporción directa a la diferencia entre las poblaciones totales de las que se hizo la selección, y en proporción inversa a la correlación entre el test y el pre-retest.

Rulon [1941], Stanley y Beeman [1958] y Thorndike [1942] han estudiado este problema en forma exhaustiva, destacando el análisis de covariancia y otras técnicas estadísticas sugeridas por Johnson y Neyman [véase Johnson y Jackson, 1959, págs. 424-44] y por Peters y Van Voorhis [1940] para probar los efectos de la variable experimental sin el procedimiento de hallar pares de grupos similares. No obstante, habría que tomar en cuenta recientes advertencias de Lord [1960] a propósito del análisis de covariancia cuando la confiabilidad de la covariable no es absoluta. También pueden aplicarse puntajes simples de ganancia, pero suelen ser menos convenientes que el análisis de covariancia. La aplicación del análisis de covariancia a esta situación del diseño 10 implica supuestos (como el de homogeneidad de regresión) menos posibles aquí que en los casos del diseño 4 [véase Lindquist, 1953].

Al interpretar estudios publicados del diseño 10, en que se recurrió a la equiparación, se puede advertir que el sentido del error es predecible. Consideremos un experimento de psicoterapia que utiliza como *O* calificaciones de descontento con la propia personalidad. Supongamos que el grupo experimental consta de personas sometidas a terapia, en tanto que el grupo de control seleccionado está formado por personas consideradas «normales». En este caso el grupo de control presentará puntajes extremadamente bajos con respecto al grupo normal (seleccionados por esta característica), y regresionará en el postest en el sentido de la media del grupo normal, haciendo así menos probable que se demuestre un efecto significativo de la terapia en vez de producir una falsa impresión de eficacia en favor del procedimiento terapéutico.

El ejemplo de los pacientes de psicoterapia nos ofrece también un caso en el cual los supuestos de regresión homogénea y muestreo del mismo universo, salvo para los puntajes extremos, parecen inapropiados. La inclusión de controles normales en la investigación psicoterapéutica es de alguna utilidad, pero hay que poner suma cautela en la interpretación de los resultados. Es importante distinguir dos versiones del diseño 10, y darles diferente jerarquía como aproximaciones a la experimentación propiamente dicha. Por una parte, se da la situación en que el experimentador dispone de dos grupos naturales, por ejemplo dos clases, y puede elegir con libertad cuál ha de recibir *X*, o por lo menos no tiene ningún motivo para sospechar que se haga un reclutamiento diferencial con relación a *X*. Aunque los grupos pueden diferir en sus medias iniciales de *O*, el estudio se aproximará a la experimentación propiamente dicha. Por otra parte, hay casos del diseño 10 en que los participantes son a todas luces autoseleccionados: el grupo experimental busca deliberadamente la exposición a *X*, y no se cuenta con un grupo de control tomado de esa misma población. En este último caso,

es menos probable que se cumpla el supuesto de regresión uniforme entre los grupos experimental y de control, aumentando en cambio la posibilidad de interacción selección-maduración (y las demás interacciones de selección). El diseño 10 «autoseleccionado» es, pues, mucho más endeble, pero no ofrece información que en muchos casos eliminaría la hipótesis de que *X* surte algún efecto. El grupo de control ayuda a interpretar, aunque sea muy divergente en el método de reclutamiento y el nivel medio.

La amenaza que la administración de tests constituye para la validez externa es la expuesta a propósito del diseño 4 (véase pág. 32). El signo de interrogación para la interacción de la selección y *X* nos recuerda que el efecto de *X* bien puede ser específico de los participantes seleccionados como lo fue de los participantes de nuestro experimento. Como los requisitos del diseño 10 pueden poner menos restricciones a nuestra libertad de muestreo que los del diseño 4, esa especificidad será por lo común menor que en un experimento de laboratorio. La amenaza a la validez externa proveniente de la reactividad de los dispositivos existe, pero tal vez en menor grado que en la mayoría de los experimentos propiamente dichos, como el diseño 4.

Donde existe la posibilidad de utilizar dos cursos intactos con el diseño 10, o la de tomar muestras aleatorias de los alumnos fuera de las aulas para distintos tratamientos experimentales según un diseño 4, 5 o 6, es casi seguro que este último dispositivo será más reactivo, creando mayor conciencia de que se está siendo sometido a experimento —la sensación de «ser un conejillo de Indias» y similares.

Los estudios de Thorndike sobre disciplina formal y transferencia [p. ej., E. L. Thorndike y Woodworth, 1901; Brolyer, Thorndike y Woodyard, 1927] constituyen otras tantas aplicaciones del diseño 10 a *X* no controladas por el experimentador. Tales estudios soslayaron, al menos en parte, el error de los efectos de regresión causados por la equiparación simple, pero habría que compararlos cuidadosamente con los métodos modernos. Así, es probable que el uso de estadísticas de covariancia produjera una prueba más contundente, por ejemplo, de trasferencia del vocabulario latino al inglés.

En otro sentido, los efectos por lo común positivos, aunque mínimos, que se hallaron podrían explicarse no como trasferencias sino como la selección en los cursos de latín de los alumnos cuyo índice anual de enriquecimiento de vocabulario habría sido mayor que el del grupo de control, aun sin la

presencia del estudio del latín. Este resultado se clasificaría aquí como interacción selección-maduración. En muchos sistemas escolares esta hipótesis rival podría verificarse ampliando la gama de las *O* previas al aprendizaje del latín que se toman en consideración como en un diseño 14.

Tales estudios constituyeron denodados esfuerzos por introducir la mentalidad experimental en la investigación de campo, y merecen que se les preste renovada atención y se los amplíe con los métodos modernos.

11. Diseños compensados

Bajo este título se reúnen todos aquellos diseños en los cuales se logra el control experimental o se aumenta la precisión aplicando a todos los participantes (o situaciones) la totalidad de los tratamientos. Esos diseños recibieron las denominaciones de «experimentos rotativos» [según McCall, 1923], «diseños compensados» [p. ej., Underwood, 1949], «diseños cruzados» [Cochran y Cox, 1957; Cox, 1958] y «diseños de conmutación» [Kempthorne, 1952]. El dispositivo de cuadrado latino es el que más se utiliza en la compensación. Ese cuadrado latino es el utilizado en el diseño 11, esquematizado aquí como cuasiexperimental, en el que se aplican cuatro tratamientos experimentales en forma restrictivamente *aleatorizada* y por turno a cuatro grupos formados de manera natural o incluso a cuatro individuos [p. ej., Maxwell, 1958]:

	Primera vez	*Segunda vez*	*Tercera vez*	*Cuarta vez*
Grupo A	X_1O	X_2O	X_3O	X_4O
Grupo B	X_2O	X_4O	X_1O	X_3O
Grupo C	X_3O	X_1O	X_4O	X_2O
Grupo D	X_4O	X_3O	X_2O	X_1O

El diseño ha sido diagramado sólo con postests, dado que presta particular utilidad allí donde los pretests resultan inapropiados y no se dispone de diseños como el 10. El diseño contiene tres clasificaciones (grupos, sesiones y *X* o tratamientos experimentales). Cada clasificación es «orto-

gonal» respecto de las otras dos, en el sentido de que cada variable de cada clasificación se produce con la misma frecuencia (una vez para un cuadrado latino) con cada variable de cada una de las otras clasificaciones. Obsérvese que cada tratamiento (o X) sólo se da una vez en cada columna y cada fila. El mismo cuadrado latino puede modificarse de tal manera que las X se conviertan en títulos de filas o de columnas:

	X_1	X_2	X_3	X_4
Grupo A	t_1O	t_2O	t_3O	t_4O
Grupo B	t_3O	t_1O	t_4O	t_2O
Grupo C	t_2O	t_4O	t_1O	t_3O
Grupo D	t_4O	t_3O	t_2O	t_1O

Resultan así comparables las sumas de puntajes por X, al tener representados, en cada una de ellas, cada oportunidad y grupo. Las diferencias en tales sumas no se podrían interpretar como resultados artificiales de las discrepancias grupales iniciales o de efectos de la práctica, la historia, etc. De parecida comparabilidad son las sumas de las filas para diferencias grupales intrínsecas, y las sumas de las columnas de la primera presentación para las diferencias en las sesiones. Desde el punto de vista del análisis de variancia, el diseño parece suministrar así información acerca de tres efectos principales con el número de casilleros que suelen exigirse para dos. Resulta evidente el costo de esta mayor eficacia: lo que parece ser un efecto principal significativo según cualquiera de los tres criterios de clasificación, acaso constituye en cambio una compleja interacción significativa entre los otros dos [Lindquist, 1953, págs. 258-64]. Las diferencias aparentes entre los efectos de las X podrían resultar un complejo efecto específico de interacción entre las diferencias grupales y las sesiones. Las inferencias sobre los efectos de X dependerán de la admisibilidad de esta hipótesis rival, y por lo tanto las estudiaremos en forma más detallada.

Digamos, en primer lugar, que la hipótesis de tal interacción es más admisible para la aplicación cuasiexperimental descrita, que para las de los cuadrados latinos en los experimentos propiamente dichos mencionados en los textos. En lo que se ha denominado la dimensión grupal, se entremezclan dos posibles fuentes de efectos sistemáticos. Ante todo, están los factores de selección sistemática implicados en la formación natural de los grupos. Cabe esperar que esos factores tengan a la vez efectos principales e interactúen con la historia, la

maduración, los efectos de la práctica, etc. Si se tuviese que organizar así un experimento con control total, cada persona debería ser asignada a cada grupo en forma independiente y aleatoria, eliminándose esta fuente tanto de los efectos principales como de la interacción, al menos en lo que concierne al error de muestreo. Es característico del cuasiexperimento que la compensación se introduzca para suministrar una suerte de igualación, solo porque tal asignación aleatoria no es posible. (Como contraste, en diseños del todo controlados, se emplea el cuadrado latino por razones de economía o para resolver problemas peculiares del muestreo de parcelas.) Una segunda posible fuente de efectos entremezclados en los grupos es la vinculada con secuencias específicas de tratamientos. Si todas las repeticiones de un experimento propiamente dicho hubiesen seguido el mismo cuadrado latino, esta fuente de efectos principales y de interacción también habría estado presente. Sin embargo, en el típico experimento propiamente dicho, a algunos grupos de participantes se les habrían asignado en la repetición diferentes cuadrados latinos, eliminándose así el efecto sistemático de secuencias específicas. De ese modo se elimina también la posibilidad de que determinada interacción sistemática haya producido un aparente efecto principal de las X.

Es probable que las sesiones produzcan un efecto principal debido a la repetida aplicación de pruebas, la maduración, la práctica y los efectos acumulados o trasferencias. Asimismo, la historia puede generar efectos con respecto a las sesiones. El dispositivo en cuadrado latino impide, por supuesto, que esos efectos principales contaminen los de X. Pero donde tales efectos son síntomas de una heterogeneidad significativa, es probable que se justifique más la sospecha de interacciones significativas que cuando tales efectos principales no se producen. Los efectos de la práctica, por ejemplo, quizá sean monótonos, pero también es probable que no sean lineales y generen efectos tanto principales como de interacción. Muchas aplicaciones de los cuadrados latinos en experimentos propiamente dichos, como en la agricultura, por ejemplo, no exigen reiteradas mediciones y es característico que no produzcan ningún efecto sistemático correspondiente de columna. Los del tipo cruzado, sin embargo, comparten este posible inconveniente con los cuasiexperimentos.

Estas consideraciones permiten apreciar la máxima importancia de la repetición del diseño cuasiexperimental con diferentes cuadrados latinos específicos. Tales repeticiones, realizadas

en número suficiente, harían del cuasiexperimento un experimento propiamente dicho. Es probable que implicasen también cantidades suficientes de grupos para posibilitar la asignación aleatoria de grupos intactos a los tratamientos, medio de control que por lo común es preferible. No obstante, careciendo de tales posibilidades, un cuadrado latino único constituye un diseño cuasiexperimental intuitivamente satisfactorio, a causa de su demostración de todos los efectos en la totalidad de los grupos de comparación. Aun reconociendo los posibles errores de interpretación, constituye un diseño que bien vale la pena adoptar cuando no hay posibilidades de un control más eficaz. Una vez destacados sus graves inconvenientes, examinemos sus ventajas relativas.

Como todos los cuasiexperimentos, gana este en pujanza con la congruencia de las repeticiones internas de la prueba. Para poner de relieve esa congruencia, deben eliminarse los efectos principales de las sesiones y los grupos, expresando cada casillero como un desvío respecto de las medias de filas (grupo) y columnas (momentos): $M_{gt} - M_{g.} - M_{.t} + M_{..}$. Después se reordenan los datos, con los tratamientos (X) encabezando las columnas. Supongamos que el cuadro que obtenemos es de una satisfactoria congruencia, que el más eficaz de los tratamientos es el mismo en los cuatro grupos, etc. ¿Cuáles son las probabilidades de que eso no sea un efecto real de los tratamientos, sino una interacción de grupos y sesiones? Podemos observar que casi todas las posibles interacciones de grupos y sesiones reducirían o enturbiarían el efecto manifiesto de X. Una interacción que imitara un efecto principal de X sería poco probable, y lo sería menos aún en cuadrados latinos mayores.

Nos sentiríamos muy atraídos por este diseño cuando tuviésemos control de programación sobre unos cuantos grupos de formación natural, como por ejemplo clases, pero no nos fuese posible subdividir esos grupos naturales en subgrupos de equivalencia aleatoria, sea para una presentación de X o para aplicar tests. En tal situación, si hubiera cómo aplicar un pretest, se dispondría asimismo del diseño 10; también implica una posible confusión de los efectos de X con interacciones de selección y sesiones. Se juzga que esta posibilidad es menos probable en el diseño compensado, porque en cada grupo se demuestran todas las comparaciones y por lo tanto se necesitarían varias interacciones equiparadas a fin de imitar el efecto experimental.

Mientras que en los otros diseños la especial sensibilidad de

uno solo de los grupos a un acontecimiento externo (historia) o a la práctica (maduración) podría simular un efecto de X, en el diseño compensado tales efectos coincidentes tendrían que darse en sucesivas sesiones separadas y en cada uno de los grupos. Este resultado supone, por supuesto, que no interpretaríamos un efecto principal de X como significativo si la inspección de los casilleros indicase que un efecto principal desde el punto de vista estadístico ha sido originado, fundamentalmente, por un muy poderoso efecto en solo uno de los grupos. Para un estudio más detenido de esta cuestión, véanse Wilk y Kempthorne [1957], Lubin [1961] y Stanley [1955].

12. Diseño de muestra separada pretest-postest

Para grandes poblaciones —p. ej., ciudades, fábricas, escuelas y unidades militares—, suele ocurrir que, aunque no se pueden segregar subgrupos en forma aleatoria para tratamientos experimentales diferenciales, cabe ejercer algo así como un control experimental completo sobre el *momento de aplicación* y los *destinatarios* de la O, utilizando procedimientos de asignación aleatoria. Ese control posibilita el diseño 12:

$$\begin{array}{llll} R & O & (X) & \\ R & & X & O \end{array}$$

En este esquema, las filas constituyen subgrupos de equivalencia aleatoria, representando la X entre paréntesis una presentación de X sin importancia. Se mide una muestra antes de X, otra equivalente después de X. El diseño no es intrínsecamente eficaz, como lo indica su fila en el cuadro 2. No obstante, suele resultar viable, y a menudo merece que se lo aplique. Se lo ha utilizado en experimentos de ciencias sociales que son aún los mejores estudios existentes en sus temas específicos [p. ej., Star y Hughes, 1950]. Aunque se lo ha denominado «diseño simulado antes-y-después» [Selltiz, Jahoda, Deutsch y Cook, 1959, pág. 116], vale la pena destacar su superioridad respecto del diseño común antes-y-después, el diseño 2, por su control tanto del efecto principal de la aplicación de tests como de la interacción de la administración de tests con X. El defecto fundamental del diseño es que no puede controlar la historia. Así, en el estudio de la campaña de publicidad realizado en Cincinnati para las Na-

ciones Unidas y la UNESCO [Star y Hughes, 1950], es probable que hechos externos de la escena internacional hubieran sido la causa de la reducción observada en el optimismo sobre la coexistencia pacífica con Rusia.

Esta obra aspira a estimular los diseños «de retazos», en los cuales se agregan aspectos que permitan controlar factores específicos, de ordinario uno por vez (en contraste con los experimentos propiamente dichos, de mayor elegancia, en que con un solo grupo se controlan todas las amenazas a la validez interna). Repitiendo el diseño 12 en diferentes situaciones y momentos, como en el diseño 12*a* (véase cuadro 2, pág. 80), se controla la historia, pues si el mismo efecto se da en varias ocasiones, la posibilidad de que sea resultado de acontecimientos históricos coincidentes se torna menos probable. No obsbante, las tendencias históricas persistentes o los ciclos estacionales siguen constituyendo explicaciones rivales no controladas. Por la repetición del efecto en otras condiciones, cabe reducir la posibilidad de que el efecto observado sea característico de la única población seleccionada en el primer momento. No obstante, si la situación de la investigación permite utilizar el diseño 12*a*, también será viable el 13, que en general resultará preferible.

Es poco probable que se invoque la maduración, o el efecto del envejecimiento de los participantes, como explicación rival, ni aun en estudios sobre la opinión pública que se extiendan durante meses. Pero en la encuesta por muestreo, y hasta en ciertos cursos universitarios, las muestras son suficientemente grandes y las edades lo bastante heterogéneas para que se puedan comparar las submuestras del grupo pretest que difieren en maduración (edad, número de semestres cursados, etc.). La maduración, y la acaso más peligrosa posibilidad de tendencias persistentes y estacionales, también es controlable por un diseño como el 12*b*, que agrega un grupo pretest anterior, aproximando el diseño al de series cronológicas, aunque sin la aplicación reiterada de tests. Para poblaciones como la de pacientes a quienes se aplican tratamientos de psicoterapia, donde podría darse una mejoría espontánea o curación, los supuestos de linealidad implicados en forma implícita en este control quizá no fueran aceptables. Es más probable que la tendencia de maduración reciba una aceleración negativa, haciendo así que la ganancia de maduración $O_1 - O_2$ sea mayor que la de $O_2 - O_3$, en detrimento, por tanto, de la interpretación de que X ha producido efecto.

La instrumentación constituye un riesgo en este diseño, cuando se la utiliza en el marco de las encuestas por muestreo. Si en el pretest y el postest se recurre a los mismos encuestadores, suele ocurrir que muchos, carentes aún de experiencia en el pretest, la hayan adquirido en el postest o tengan en él mayor soltura. Si en cada tanda de encuestas se recurre a distintas personas para esa tarea, y su número no es elevado, las diferencias en la idiosincrasia de los encuestadores se confunden con la variable experimental. Si los experimentadores conocen la hipótesis, sus expectativas pueden provocar diferencias, háyase o no trasmitido la X, como lo demostraron con sus experimentos Stanton y Baker [1942] y Smith y Hyman [1950]. En un caso ideal se utilizarían muestras aleatorias equivalentes de distintos entrevistadores en cada tanda, manteniéndolos ignorantes acerca del objeto del experimento. Además, el reclutamiento de los encuestadores puede indicar diferencias estacionales, por ejemplo, ya que durante los meses de verano se dispone de más estudiantes universitarios, etc. Las tasas de rechazo son acaso menores y la duración de las entrevistas mayor en verano que en invierno. Para cuestionarios autoadministrados en el aula, este error instrumental será menos probable, aunque las orientaciones hacia la administración de tests quizá se desplacen en formas mejor clasificables como instrumentación que como influjos de X sobre O.

Para pretests y postests aplicados con varios meses de separación, la mortalidad puede plantear un problema en el diseño 12. Si ambas muestras se eligen en forma simultánea (punto R), es de suponer que a medida que trascurra el tiempo más integrantes de la muestra elegida se tornen inaccesibles, perdiéndose los segmentos más transitorios de la población, lo cual producirá una diferencia poblacional entre los distintos períodos de entrevista. Una advertencia de esa posibilidad la constituyen las diferencias entre los grupos en el número de personas no entrevistadas.

En estudios realizados a lo largo de períodos extensos, las muestras para pretest y postest deberían seleccionarse acaso en forma independiente y en momentos distintos apropiados, aunque ello también posee una fuente de sesgo sistemático, resultante de los posibles cambios en el esquema residencial del conjunto del universo. En algunos medios (p. ej., en las escuelas, los archivos permitirán que se eliminen los puntajes pretest de quienes no estarán ya disponibles en el momento del postest, haciendo así más comparables el postest con el pretest. Para lograr un mecanismo que haga posible esa corrección en la encuesta con muestras, así como una ratificación

del efecto que no pudiera contaminarse con la mortalidad, se puede someter el grupo pretest a un nuevo test, como en el diseño 12*c*, donde la diferencia $O_1 - O_2$ confirmaría la comparación $O_1 - O_3$. Así, el estudio que Duncan y otros [1957] efectuaron sobre la reducción en las creencias erróneas lograda durante un curso introductorio de psicología. (En este diseño, el grupo sometido a un retest no permite que se examinen las ganancias de personas con puntajes iniciales diversos, por no haberse utilizado un grupo de control para verificar la existencia de regresión.)

Lo característico de este diseño es que lleva el laboratorio a la situación de campo a la cual el investigador desea extender sus generalizaciones, probando los efectos de *X* en su ambiente natural. En general, según se indica en los cuadros 1 y 2, los diseños 12, 12*a*, 12*b* y 12*c* pueden resultar superiores en validez externa o posibilidad de generalización respecto de los experimentos propiamente dichos de los diseños 4, 5 y 6. Estos diseños no requieren gran cooperación de los participantes, ni que estén disponibles en ciertos lugares y momentos, etc., de modo que se puede utilizar un muestreo representativo de poblaciones previamente determinadas.

En los diseños 12 y 13 (y sin lugar a dudas también en algunas variantes de los diseños 4 y 6, donde *X* y *O* se trasmiten por contactos individuales, etc.), es posible el muestreo representativo. Los signos positivos en la columna de interacción selección-*X* son muy relativos y con todo derecho se los podría cambiar por signos de interrogación ya que en la práctica general las unidades no se seleccionan por su relevancia teórica, sino a menudo por razones de cooperación y accesibilidad, que posiblemente las tornen atípicas del universo al cual se las desea generalizar.

Star y Hughes [1950] no deseaban generalizar a Cincinnati, sino más bien a los ciudadanos de Estados Unidos o al mundo en general, y persiste la posibilidad de que la reacción a *X* en aquella urbe fuese atípica de esos universos. Pero el grado de ese sesgo de accesibilidad es tan inferior al de otros diseños más exigentes que, en comparación, parece justificado atribuirle un carácter positivo.

13. Diseño de muestra separada pretest-postest con grupo de control

Se supone que el diseño 12 ha de utilizarse en aquellas situaciones en que la *X*, si existe, debe presentarse al grupo como un todo. Si se cuenta con grupos comparables (ya que no equivalentes) a los cuales sea posible rehusar la *X*, se podrá agregar un grupo de control al diseño 12, creando así el diseño 13:

R	*O*	(*X*)			
R			*X*	*O*	
R	*O*				
R					*O*

Este diseño es bastante parecido al 10, solo que no se vuelve a someter a test a las mismas personas y, por lo tanto, se evita la posible interacción entre la administración de tests y *X*. Como en el diseño 10, la desventaja del 13 en cuanto a la validez interna proviene de la posibilidad de interpretar como efecto de *X* otra tendencia local propia del grupo experimental que, en realidad, no ha influido. Aumentando el número de las unidades sociales implicadas (escuelas, ciudades, fábricas, buques, etc.) y asignándolas en cierto número y con aleatorización a los tratamientos experimentales y de control, se conseguirá eliminar la única fuente de invalidación, lográndose así un experimento propiamente dicho, análogo al diseño 4, con la única diferencia de que se evitan nuevas pruebas sobre los mismos individuos. Este diseño puede designarse 13*a*. Su esquematización (en el cuadro 3) se ha visto complicada por los dos niveles de equivalencia (logrados por asignación aleatoria) en él implicados. En el nivel de participantes, existe en el interior de cada unidad social la equivalencia de las muestras separadas pretest y postest, indicadas por el punto *R* de asignación. Entre las varias unidades sociales que reciben cualquiera de los tratamientos, no se verifica esa equivalencia, lo cual se indica con la línea punteada. La *R'* designa la igualación del grupo experimental y el de control por la asignación aleatoria de esas muchas unidades sociales a uno u otro tratamiento.

Como puede verse en la fila correspondiente a 13*a* del cuadro 3, este diseño obtiene un puntaje perfecto para validez

tanto interna como externa, esta última en virtud de los fundamentos ya expuestos a propósito del diseño 12, y con mayor hincapié en el problema de la interacción selección-*X*, a causa de que están representadas muchas unidades sociales y no una sola. Que nosotros sepamos, este diseño, excelente pero costoso, no ha sido utilizado nunca.

14. Diseño de series cronológicas múltiples

En los estudios de grandes cambios administrativos por medio de datos en series cronológicas, al investigador le conviene buscar una institución similar no sujeta a *X*, de la cual tomar una serie cronológica de «control» análoga (idealmente, con *X* asignada al azar):

$$\frac{O\ O\ O\ OXO\ O\ O\ O}{O\ O\ O\ O\ O\ O\ O\ O}$$

Este diseño contiene (en las *O* que comprenden a *X*) el número 10, de grupo de control no equivalente, pero gana certidumbre de interpretación por las múltiples mediciones representadas, ya que en cierto sentido el efecto experimental se demuestra dos veces, respecto del control y respecto de los valores pre-*X* en su propia serie, como en el diseño 7. Además, la interacción entre selección y maduración se controla en el sentido de que, si el grupo experimental demostró por lo común una mayor tasa de ganancia, aparecería así en las *O* pre-*X*. En los cuadros 2 y 3 es escasa la representación de esta nueva ganancia, pero aparece en la columna final de validez interna, titulada «Interacción de selección y maduración». Puesto que la maduración se controla tanto en la serie experimental como en la de control, por las razones expuestas en nuestra primera presentación del diseño 7 de serie cronológica, la diferencia en la selección de los grupos, que opera juntamente con la maduración, instrumentación o regresión, difícilmente podrá explicar un efecto notorio. Sin embargo, no se excluye la posibilidad de una interacción entre la diferencia de selección y la historia.
Como con el diseño 7 de serie cronológica, se ha puesto un signo negativo en la columna de validez externa para la in-

Cuadro 3. *Fuentes de invalidación para los diseños 13 a 16.*

	Fuentes de invalidación										
	Interna								Externa		
	Historia	Maduración	Administración de tests	Instrumentación	Regresión	Selección	Mortalidad	Interacción de selección y maduración, etc.	Interacción de administración de tests y *X*	Interacción de selección y *X*	Dispositivos reactivos Interferencia de *X* múltiples
Diseños cuasiexperimentales (cont.)											
13. Diseño de muestra separada pretest-postest con grupo de control *R O (X)* *R X O* *R O* *R O*	+	+	+	+	+	+	+	−	+	+	+
13*a* *R'* *R O (X)* *R X O* *R O (X)* *R X O* *R'* *R O* *R O* *R O* *R O*	+	+	+	+	+	+	+	+	+	+	+
14. Series cronológicas múltiples *O O O XO O O* *O O O O O O O*	+	+	+	+	+	+	+	+	−	−	?
15. Diseño de ciclo institucional Cl. *A X O*$_1$ Cl. *B*$_1$ *RO*$_2$ *X O*$_3$ Cl. *B*$_2$ *R X O*$_4$ Cl. *C O*$_7$ *X* Cont. Gen. Pob. p/Cl. *B O*$_6$ Cont. Gen. Pob. p/Cl. *C O*$_7$											
$O_2 < O_1$ $O_5 < O_4$	+	−	+	+	?	−	?		+	?	+
$O_2 < O_3$	−	−	−	?	?	+	+		−	?	+
$O_2 < O_4$	−	−	+	?	?	+	?		+	?	?
$O_6 = O_7$ $O_{2y} = O_{2o}$		+						−			
16. Discontinuidad en la regresión	+	+	+	?	+	+	?	+	+	−	+ +

«Cont. Gen. Pob. p/Cl.» significa «Controles generales de población para la clase».

teracción entre la aplicación de pruebas y *X*, aunque como en el caso del mismo diseño 7, el que comentamos se empleará a menudo cuando la administración de los tests no sea reactiva. También la habitual preocupación acerca de la posible especificidad de un efecto demostrado de *X* en la población que se estudia queda registrada en el cuadro 3. En cuanto a los tests de significación, se sugiere que las diferencias entre la serie experimental y la de control se analicen como los datos del diseño 7. Parece mucho más probable la linealidad de estas diferencias que la de los datos no elaborados de las series cronológicas.

Este es, en términos generales, un excelente diseño cuasiexperimental, acaso el mejor de los más viables. Presenta claras ventajas respecto de los diseños 7 y 10, como ya lo hemos indicado al presentar el diseño 10. La posibilidad de efectuar reiteradas mediciones torna particularmente apropiadas las series cronológicas múltiples para las investigaciones que se llevan a cabo en establecimientos educativos.

15. Diseño de ciclo institucional recurrente: un diseño «de retazos»

El diseño 15 ilustra una estrategia para la investigación de campo en la cual se comienza por un diseño insuficiente y se van sumando luego características particulares, a fin de investigar una u otra de las fuentes recurrentes de invalidación. El resultado es a menudo una burda acumulación de verificaciones precautorias, que carece de la simetría intrínseca de los diseños experimentales propiamente dichos, pero se asemeja a la experimentación. Como parte de esa estrategia, el experimentador habrá de estar alerta ante las interpretaciones antagónicas (ajenas al efecto de *X*) que el diseño ofrece, y tendrá que buscar la explicación de los datos, o las posibles extensiones de ellos, que permitirían descartarlas. Otro aspecto bastante característico de estos diseños es que el efecto de *X* se demuestra en varias formas diferentes. Este aspecto es importante sin duda cuando cada comparación específica sea equívoca por sí sola.

El diseño «de retazos» específico que exponemos se limita a un riguroso conjunto de cuestiones y situaciones, y explota según las circunstancias las características que estas exhiben.

La idea fundamental puede apreciarse en las filas segunda y tercera del cuadro 1, donde se advierte que los signos positivos y negativos de los diseños 2 y 3 son en su mayor parte complementarios, y que, en consecuencia, la correcta combinación de esos dos criterios, insuficientes por sí solos, podría tener gran vigor. El diseño es apropiado para aquellas situaciones en que se presenta en forma cíclica, a cada nuevo grupo de participantes, cierto aspecto de un proceso institucional (escuelas, métodos de adoctrinamiento, aprendizaje de oficios, etc.). Si en esas situaciones nos interesa la evaluación de los efectos de una *X* tan global y compleja como un programa de adoctrinamiento, es probable que el diseño de ciclo institucional recurrente ofrezca la respuesta más aproximada posible resultante de los diseños que hasta aquí hemos expuesto.

El diseño se ideó originariamente durante una investigación de los efectos de un año de entrenamiento para oficiales y pilotos sobre las actitudes hacia los superiores y los subordinados y las funciones de liderazgo de un grupo de cadetes de la Fuerza Aérea, mientras se completaba un ciclo de entrenamiento de 14 meses [Campbell y McCormack, 1957]. La restricción que impidió que se realizara un experimento propiamente dicho fue la imposibilidad de controlar quiénes estarían expuestos a la variable experimental. No había forma de dividir el curso de ingreso en dos mitades igualadas, una de las cuales cursaría el programa anual planificado, mientras que a la otra se la haría volver a la vida civil. Aun en el supuesto de que fuese posible un experimento propiamente dicho de esa índole (y el aprovechamiento oportuno de imprevistas reducciones presupuestarias pudo haberlo hecho posible en más de una ocasión), los efectos reactivos de ese dispositivo experimental —el inevitable trastorno en las vidas de quienes fuesen aceptados, seleccionados, trasportados a la base aérea y devueltos después a sus casas— distaría mucho de hacer de ellos un grupo ideal de control. La diferencia entre ellos y el grupo experimental que recibiría el adoctrinamiento difícilmente podría constituir una base adecuada sobre la cual generalizar las conclusiones obtenidas a las condiciones normales de reclutamiento y entrenamiento de las milicias. Quedaba, sin embargo, el control del experimentador sobre la programación del *momento* y los *destinatarios* de los procedimientos de observación. Esto, más el hecho de que la variable experimental era recurrente y se presentaba constantemente a cada nuevo grupo de participantes, hizo posible cierta forma

de control experimental. En aquel estudio se disponía de dos clases de comparaciones relativas al influjo de la experiencia militar sobre las actitudes. Cada una de ellas era bastante insuficiente desde el punto de vista del control experimental, pero cuando ambas suministraron pruebas coincidentes, se ratificaron entre sí en la medida en que ambas incluían sus respectivos puntos débiles. La primera ofrecía comparaciones entre poblaciones medidas al mismo tiempo pero con distinta duración de servicio. La segunda incluía mediciones del mismo grupo de personas en su primera semana de entrenamiento militar y otra vez después, trascurridos ya unos 13 meses de servicio. Un tanto estilizado, el diseño es como sigue:

Clase A	X	O_1
Clase B	O_2 X	O_3

Este diseño combina los enfoques «longitudinal» y de «corte trasversal» que suelen emplearse en la investigación del desarrollo. En esta se supone que la comparación es tal, que pueden medirse a la vez un grupo expuesto a X y otro que va a serlo; esta comparación entre O_1 y O_2 corresponde así al diseño 3, «Comparación de grupos estáticos». La segunda medición del personal de la Clase B, un ciclo después, nos da el segmento de diseño 2, «Pretest-postest de un grupo». En el cuadro, pág. 109, las dos primeras filas referentes al diseño 15 muestran un análisis de esas comparaciones. La comparación cruzada de $O_1 > O_2$ suministra diferencias que no podrían explicarse por los efectos de la historia o por el test-retest, sino que podrían deberse a diferencias en el reclutamiento de un año a otro (como se indica por medio del signo negativo en «Selección») o a la circunstancia de que los participantes eran un año mayores (signo negativo en «Maduración»). Cuando todas las pruebas se realizan durante el mismo período, parece improbable que haya una variable entremezclada de instrumentación o desvíos en la índole del instrumento de medición. En la típica comparación de las diferencias de actitud entre alumnos universitarios de primero y segundo año, el efecto de la mortalidad no pasa de ser una explicación rival: O_1 y O_2 podrían diferir solo a causa del tipo de personas que han abandonado sus estudios en la Clase A, pero continúa teniendo representación en la B. Este inconveniente se puede evitar si las reacciones se identifican por individuos y el experimentador espera antes de analizar sus datos a que la Clase B haya

completado su exposición a X y luego elimina de O_2 todas las medidas pertenecientes a participantes que después no completaron su instrucción. La frecuente ausencia de este procedimiento justifica la inserción de un signo interrogativo al lado de la variable de mortalidad. La columna «Regresión» se completa con signos interrogativos a fin de señalar la posibilidad de efectos espurios si la medida que se utiliza en el diseño experimental es la misma en que se fundan la aceptación o el rechazo de candidatos al curso de entrenamiento. En tales circunstancias serían de prever diferencias constantes no atribuibles a los efectos de X. La comparación pretest-postest implicada en O_2 y O_3, si resulta ser el mismo tipo de diferencia que en la comparación $O_2 - O_1$, elimina las demás hipótesis posibles de que la diferencia se deba a un desvío en la selección o reclutamiento entre ambas clases, así como cualquier posibilidad de que la mortalidad haya sido la causa. No obstante, si no se utilizara más que la comparación $O_2 - O_3$, sería vulnerable a las explicaciones rivales de historia y aplicación de tests.

En una situación donde el lapso de entrenamiento que se examina es de un año, el aspecto más costoso del diseño es la programación de ambos conjuntos de mediciones con un año de diferencia. Dada la inversión ya realizada en este sentido, constituye un pequeño gasto más realizar nuevas pruebas en la segunda ocasión. Teniendo en cuenta todo ello, cabe extender el diseño institucional recurrente al esquema indicado en el cuadro 3. Ejerciendo el poder de designar cuándo y a quién se ha de medir, la Clase B se ha dividido en dos muestras igualadas, una medida antes y después de la exposición y la otra medida solo después de ella, como en O_4. Este segundo grupo permite una comparación, sobre muestras cuidadosamente igualadas, de una medición inicial «antes y después»; es más precisa que la comparación $O_1 - O_2$ en lo que respecta a la selección, y superior a la comparación $O_2 - O_3$, ya que evita los efectos de test-retest. El efecto de X queda así documentado por medio de tres comparaciones distintas, $O_1 > O_2$, $O_2 < O_3$ y $O_2 < O_4$.

Nótese, sin embargo, que O_2 aparece en las tres, razón por la cual todo ello podría parecer confirmatorio solo en virtud de una actuación excéntrica del mencionado conjunto particular de mediciones. La introducción de O_5, o sea la Clase C, probada en ocasión del segundo test antes de ser expuesta a X, ofrece una nueva medición pre-X que puede compararse con O_4 y O_1, etc., brindando una redundancia necesaria. La divi-

sión de la Clase B hace esta comparación de $O_4 - O_5$ más clara que lo que sería una $O_3 - O_5$. Adviértase, empero, que la división de una clase en dos mitades, sometida una a test y la otra no, suele constituir un dispositivo reactivo. Por eso se ha incluido un signo de interrogación para ese factor en la fila $O_2 < O_4$ del cuadro 3. Que sea o no un procedimiento reactivo depende de las condiciones concretas. Cuando se echan suertes y se pide que la mitad de la clase pase a otra aula, es probable que el procedimiento sea reactivo [p. ej., Duncan y otros, 1957; Solomon, 1949]. Cuando, como sucede en muchos estudios sobre militares, las entrevistas se han realizado en forma individual, una clase puede dividirse en mitades iguales sin que el hecho resulte tan ostensible. Cuando un curso está formado por un cierto número de divisiones con programas diferentes, hay la posibilidad de asignar esas unidades intactas a los grupos con pretest y sin él [p. ej., Hovland, Lumsdaine y Sheffield, 1949]. Para una clase única, el recurso de distribuir cuestionarios o tests a todos, pero variando el contenido a fin de que una mitad aleatoria obtenga lo que constituiría el pretest y la otra se pruebe con algún otro instrumento, puede servir para lograr que la división del curso no sea más reactiva que el test de la clase total.

El diseño, tal como se lo representa por medio de las mediciones O_1 a O_5 falla siempre en el control de la maduración. La gravedad de esa limitación variará de acuerdo con el material que se investigue. Si el experimento versa sobre la adquisición de una habilidad o técnica muy poco común, la hipótesis rival de maduración —que el simple hecho de envejecer o de adquirir experiencia gracias a las prácticas sociales cotidianas habría producido esa habilidad— puede resultar sumamente improbable.

Sin embargo, en el citado estudio de actitudes hacia superiores y subalternos [Campbell y McCormack, 1957], el desvío fue tal que bien podría explicarse a causa de la mayor preparación que, casi en cualquier contexto, habría adquirido un grupo de aquella edad y tipo particular de ambiente al crecer en edad o estar lejos de sus hogares respectivos. En tal situación parece imprescindible un control de maduración. Por ese motivo se han agregado O_6 y O_7 al diseño, a fin de ofrecer una prueba de corte trasversal de una hipótesis general de maduración hecha en ocasión del segundo período de tests. Ello exigirá someter a prueba a dos grupos de personas de la población general que solo difieran en la edad, la cual se elegiría a fin de que coincidiera con las de las Clases B y C en

la época de las pruebas. Para confirmar la hipótesis de un efecto de X, los grupos O_6 y O_7 deberían ser iguales, o al menos acusar una discrepancia menor que las comparaciones que abarcan la exposición a X. La selección de tales controles poblacionales generales dependería de lo específico de la hipótesis. Dado nuestro conocimiento acerca de la universal importancia de las consideraciones de clase social y educación, esos controles podrían seleccionarse de tal modo que equiparasen el reclutamiento institucional con la clase social y la educación anterior. Asimismo, podrían ser personas que vivieran fuera de sus hogares por primera vez y que tuviesen la edad típica de búsqueda de independencia; así en el ejemplo dado, el grupo O_6 habría estado lejos de su casa durante un año, y el O_7 estaría a punto de abandonarla. Esos controles de relación de edad en la población general serían siempre hasta cierto punto insatisfactorios y constituirían el rubro más costoso, ya que la aplicación de pruebas dentro del esquema de una institución es por lo común más simple que seleccionar casos de una población general. Por esa razón, O_6 y O_7 han sido programados con la segunda tanda de pruebas, pero si no resulta ningún efecto de X en el primer conjunto de resultados (la comparación $O_1 > O_2$), tan costosos procedimientos estarían por lo común injustificados (a menos, claro está, que se propugnase la hipótesis de que la X institucional había eliminado un proceso normal de maduración). Otro enfoque por corte trasversal del control de la maduración puede darse si hay heterogeneidad de edades (o un cierto número de años fuera del hogar, etc.) dentro de la población que ingresa en el ciclo institucional. Así ocurriría en muchas situaciones; por ejemplo, al estudiar los efectos de un curso universitario aislado. En este caso, las mediciones de O_2 podrían subdividirse en un grupo de mayor y otro de menor edad, a fin de examinar si esos dos subgrupos (O_{2o} y O_{2y} en el cuadro 3) diferían como lo habían hecho O_1 y O_2 (aunque la universal correlación negativa entre edad y capacidad *dentro* de los grados escolares, etc., introduce aquí no pocos peligros). Mejor que el control con los coetáneos de toda la población, la comparación podría hacerse con otra institución determinada, por ejemplo, entre los conscriptos de la Fuerza Aérea y los estudiantes universitarios de primer año. Si se ha de hacer una comparación de esta índole, se reduce la variable experimental a aquellos aspectos que ambas instituciones *no tienen* en común. En tal caso, es probable que los diseños 10 y 13, por lo común más eficaces, sean igualmente factibles.

Los requisitos formales de este diseño parecen aplicables incluso a un problema como el de la psicoterapia. Esta posibilidad revela cuán difícil es una verificación correcta de la variable maduración. Comoquiera que se elijan los controles poblacionales para una situación de psicoterapia, si no reciben este tipo de tratamiento diferirán en aspectos importantes. Aunque estén tan enfermos como los sometidos a tratamiento psicoterapéutico, es casi seguro que diferirán en su conocimiento de él, así como en sus creencias al respecto y su fe en ese procedimiento curativo. Un grupo de esta índole, enfermo pero optimista, podría muy bien tener posibilidades de recuperación típicas de cualquier grupo de comparación de que pudiésemos echar mano y, por consiguiente, podría malinterpretarse una interacción de selección y maduración como un efecto de X.

Para el estudio aislado de procesos de desarrollo, el no poder controlar la maduración no es, ciertamente, un inconveniente, ya que ella es el objetivo mismo del análisis. Esa combinación de comparaciones longitudinales y de corte trasversal debería emplearse en forma más sistemática en este tipo de estudios. El estudio aislado de cortes trasversales confunde maduración con selección y mortalidad. El estudio longitudinal confunde maduración con aplicación reiterada de tests e historia. Por sí solo no es probablemente mejor que el de corte trasversal, aunque su costo más elevado le otorga mayor prestigio. La combinación de ambos, quizá con reiteradas comparaciones de cortes trasversales en diversos momentos, parece ideal.

Tal como se presentan los esquemas del diseño 15, se supone que se podrá aplicar el postest a un grupo al mismo tiempo que el pretest a otro. No siempre ocurre así en situaciones en que tal vez se deseara utilizar este diseño. La siguiente es una representación más precisa del caso típico en la situación escolar:

Clase A	X	O_1			
Clase B_1			RO_2	X	O_3
Clase B_2			R	X	O_4
Clase C					O_5 X

Este diseño carece del claro control sobre la historia en las comparaciones $O_1 > O_2$ y $O_4 > O_5$, por falta de simultaneidad. No obstante, difícilmente podría aceptarse la explicación desde el punto de vista de la historia si ambas comparaciones acusaran el efecto, como no fuera postulando una serie bastante compleja de coincidencias.

Nótese que ninguna tendencia histórica general, como la que sin duda hallamos en las actitudes sociales, se confunde con resultados experimentales concretos. Una tendencia de esa índole colocaría a O_2 en posición intermedia entre O_1 y O_3, mientras que la hipótesis de que X tiene un efecto exige que O_1 y O_3 sean iguales y O_2 difiera de ambas en el mismo sentido. En general, si se repite varias veces el experimento, es poco probable que la confusión con la historia constituya un problema, ni siquiera en esta versión del diseño. Pero, para ciclos institucionales de menos de un año, habrá posibilidad de confusión con variaciones estacionales en actitudes, moral, optimismo, inteligencia, etc. Si la X es un curso desarrollado solo en la temporada de otoño,* y entre setiembre y enero la gente suele experimentar mayor agresividad y pesimismo a causa de los factores climáticos de la estación, esa tendencia estacional recurrente se confundirá con los efectos de X en todas sus manifestaciones. Para situaciones de esta índole pueden utilizarse, y resultan aconsejables, los diseños 10 y 13.

Si las comparaciones de corte trasversal y longitudinales indican efectos análogos de X, ello sería inexplicable como interacción entre la maduración y las diferencias de selección entre las clases. No obstante, se ha dejado en blanco la columna porque este control no aparece en las presentaciones fragmentarias del cuadro 3. Las calificaciones de los criterios de validez externa se ajustan en general al esquema de los diseños anteriores que contienen los mismos fragmentos. Los signos de interrogación en la columna «Interacción de selección y X», advierten simplemente que los descubrimientos se limitan al ciclo institucional que se estudia. Dada la complejidad de X, es posible que se realice la investigación por razones prácticas más que con propósitos teóricos, y tal vez se quiera en este caso generalizar a una institución en particular.

* Vale decir, la primera en el hemisferio Sur. (*N. del E.*)

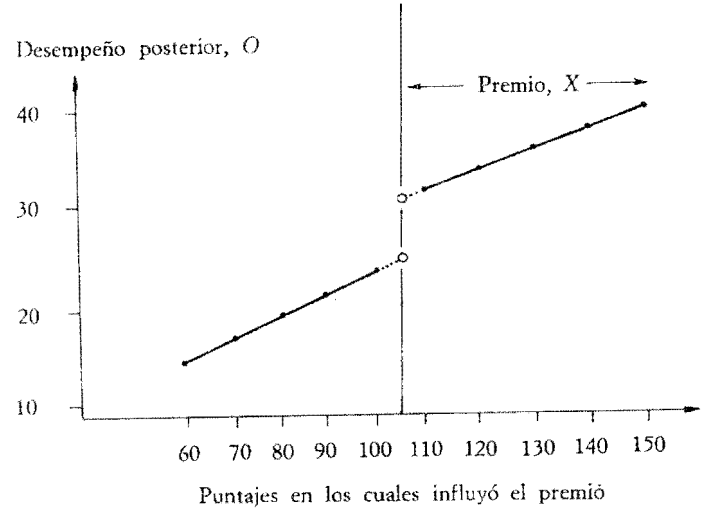
16. Análisis de discontinuidad en la regresión

Este diseño es practicable en una situación en que se han utilizado ya diseños *ex post facto*. Aunque de muy limitada aplicación, parece justificado presentarlo aquí por el hecho de que esas situaciones poco numerosas son, en su mayor parte, educacionales. También parece oportuno incluirlo como ejemplo de la conveniencia de indagar, en cada situación concreta, todas las implicaciones de una hipótesis causal, buscando nuevos afloramientos de esta última, mediante los cuales se la pudiera verificar. La situación que tomaremos [Thistlethwaite y Campbell 1960] consiste en el otorgamiento de premios a los aspirantes más calificados, sobre la base de un puntaje de corte dentro de un conjunto cuantificado de calificaciones. El premio puede ser una beca, el ingreso en una universidad tan prestigiosa que todos los aprobados se inscriben en ella, un año de estudios en Europa, etc. Después de ese acontecimiento, tanto los solicitantes que reciben el premio como los que no lo obtienen son objetos de mediciones respecto de varias *O* que representan logros, actitudes, etc., posteriores. Se plantea entonces el interrogante de si el premio provoca alguna diferencia. El problema de inferencia es difícil porque casi todas las cualidades que acreditan a un alumno para el premio (salvo, a veces, otros factores, como sus necesidades económicas y el estado en que reside) son las mismas que habrían llevado a un mejor desempeño en esas *O*. Tenemos casi la certeza anticipada de que los premiados habrían obtenido puntajes superiores en las *O* que quienes no lo fueron, aunque no se hubiesen otorgado los premios.

La figura 4 presenta el tema del diseño. Ilustra la relación prevista entre capacidad pre-premio y rendimientos posteriores, más los resultados adicionales de las oportunidades educacionales o motivacionales consiguientes. Consideremos ante todo un experimento propiamente dicho del tipo del diseño 6, con el cual contrastaremos nuestro cuasiexperimento. Ese experimento propiamente dicho podría racionalizarse como un proceso de solución de empate, o como un experimento adicional, en el que, para una estrecha amplitud de puntajes en el punto de corte o por debajo pero muy cerca de él, la asignación aleatoria daría lugar a un grupo experimental ganador del premio y un grupo de control no ganador. Es de presumir que tales grupos tendrían un desempeño similar al representado por los dos círculos en la línea de corte de la figura 4. Para esa estrecha amplitud de capacidades, se lograría un experimento propiamente dicho. *Tales experimentos son factibles y habría que realizarlos.*

El diseño cuasiexperimental 16 trata de establecer ese experimento propiamente dicho examinando la línea de regresión para una discontinuidad en el punto de corte, claramente implícita en la hipótesis causal. Si el resultado fuese como el diagramado y los círculos de la figura 4 representasen extrapolaciones de las dos mitades de la línea de regresión, y no un experimento de solución de empate dividido al azar, la prueba del efecto sería casi tan incontestable como en el experimento propiamente dicho.

Figura 4. *Análisis de discontinuidad en la regresión.*



Algunos de los tests de significación estudiados en el diseño 7 son también aplicables aquí. Nótese que la hipótesis es a todas luces de diferencia de ordenada más que de pendiente, y que el paso tiene que estar localizado en el punto *X* de la línea de regresión: cualquier «desfasaje» o «dispersión» es incompatible con la hipótesis. Son, pues, apropiadas las pruebas paramétricas y no paramétricas que evitan supuestos de linealidad. Nótese asimismo que tales supuestos son por lo común más aceptables para los datos de regresión que para series cronológicas. (Con determinados tipos de datos, como

los porcentajes, puede ser necesaria una trasformación lineal). Tal vez sea conveniente efectuar una prueba *t* vinculada con la diferencia entre los dos puntos linealmente extrapolados. Acaso el test más eficaz fuera un análisis de covariancia, en el cual el puntaje de decisión de otorgamiento del premio sería la covariable de los rendimientos ulteriores, y el tratamiento estaría representado por la adjudicación o no adjudicación del premio.

¿Es probable la aplicación de este tipo de diseño? Sin duda alguna se refiere a una situación recurrente en la cual abundan las afirmaciones en favor de la eficacia de *X*. ¿Vale la pena verificar esas afirmaciones? Un sacrificio necesario es que todos los elementos que entran en la decisión final se combinen en un índice compuesto, determinando con nitidez el punto de corte. Pero estamos convencidos de que todos los factores que influyen en una decisión —el aspecto que presenta la fotografía, la jerarquía del curso deducida de la reputación de la escuela secundaria, las relaciones del padre con los directivos del establecimiento, etc.—, pueden incluirse en un índice de esta índole, por medio de puntajes, si no se cuenta con un medio más directo. También deberíamos estar ya convencidos [Meehl, 1954] de que una fórmula de ponderación correlacional múltiple para la combinación de los elementos (aun empleando como criterio decisiones anteriores del comité de selección) suele ser mejor que las ponderaciones de un comité en cada caso particular. Nada perderíamos, pues, y mucho se podría ganar en todo sentido, cuantificando las decisiones de todo tipo relativas al premio. De proceder así, y si se llevasen registros de otorgamientos y rechazos, cabría hacer un seguimiento de los efectos varios años después.

Acaso convenga relatar aquí una parábola verídica. Una generosa fundación, interesada en mejorar la educación superior, donó a una universidad de Estados Unidos medio millón de dólares para que estudiase los efectos de la escuela sobre sus alumnos. Diez años después no había aparecido un solo informe ni siquiera remotamente relacionado con el tema. ¿Tomaron con alguna seriedad los donantes o los favorecidos con la donación las especificaciones de la propuesta formal? ¿Existía alguna respuesta posible al interrogante propuesto? Los diseños 15 y 16 parecen ofrecer las únicas aproximaciones viables. Pero, por supuesto, quizá ningún científico sienta verdadera curiosidad por los efectos de una *X* tan global.

Veamos el análisis del cuadro 3. A causa de la sincronía entre

el grupo experimental y el de control, historia y maduración parecen estar controladas. La administración de tests como efecto principal también lo está, pues tanto el grupo experimental como el de control la han recibido. Los errores de instrumentación bien podrían plantear un problema si la *O* de seguimiento se realizase bajos los auspicios determinantes del premio, en el sentido de que la gratitud por haberlo recibido y el resentimiento por lo contrario pudieran inducir a distintas manifestaciones de actitud, mayor o menor exageración del éxito alcanzado en la vida, etc. Este defecto se observaría asimismo en el experimento propiamente dicho de solución de empate. Podría controlárselo haciendo que los seguimientos los efectuara otro organismo o institución diferente. Creemos, conforme a los argumentos que anteceden, que tanto la regresión como la selección están controladas en lo que atañe a sus posibles contribuciones espurias a la inferencia, aun cuando la selección sea sesgada y exista regresión: ambas han sido controladas al representárselas en detalle, no mediante la igualación. La mortalidad constituiría un problema si fuese el ente de otorgamiento del premio el que dirigiese la medición de seguimiento, pues quienes lo recibieron, los ex alumnos, etc., cooperarían probablemente con mucha mejor disposición que los no ganadores. Nótese cómo el deseo, por lo común conveniente, del investigador de lograr que la muestra seleccionada sea bien representativa puede inducir aquí a error. Si la conducción del seguimiento con un membrete distinto provocara una reducción en la cooperación, por ejemplo, del 90 al 50 %, el experimentador tal vez se resistiese a hacer el cambio, ya que él tiene como meta una representación del 100 % de los ganadores del premio. Es posible que olvide que su verdadero objetivo es obtener datos interpretables, que ningún dato es interpretable si está aislado, y que es imprescindible contar con un grupo de contraste similar a fin de utilizar la información que posee sobre los ganadores. Tanto por este motivo como por el problema de instrumentación, quizá fuese mejor desde el punto de vista científico tener auspicios independientes y un 50 % de respuestas de ambos grupos en vez de un 90 % de los ganadores y un 50 % de los no ganadores. Una vez más, el problema de la mortalidad sería el mismo para el experimento propiamente dicho de solución de empate. En ambos casos, la amenaza que implica para la validez interna la interacción selección-maduración queda controlada. En cuanto al cuasiexperimento, se lo controla en el sentido de que esa

interacción no podría dar una explicación lícita de una clara discontinuidad de la línea de regresión en *X*. La amenaza a la validez externa de una interacción aplicación de pruebas-*X* queda controlada, a su vez, en la medida en que las mediciones fundamentales utilizadas al decidir la adjudicación del premio integran el universo al cual se quiere generalizar. Tanto el verdadero experimento de solución de empate como el análisis de discontinuidad de regresión están particularmente sujetos a la limitación de la validez externa por la interacción selección-*X*, ya que el efecto sólo ha quedado demostrado con respecto a una gama muy estrecha de talentos, es decir, sólo para los situados en el puntaje de corte. En el cuasiexperimento, las posibilidades de inferencia tal vez parezcan más amplias, pero nótese que los inconvenientes del supuesto de ajuste lineal son mínimos cuando se los extrapola solamente a un punto, como en el diseño ilustrado en la figura 4. Generalizaciones más amplias implican la extrapolación del ajuste debajo de *X* a través de toda la gama de valores de *X*, y en cada grado mayor de extrapolación aumenta el número de hipótesis rivales aceptables. También aumenta la dispersión de los valores extrapolados de diferentes tipos de curvas ajustadas a los valores inferiores a *X*, etcétera.

6. Diseños correlacionales y «ex post facto»

Una de las dimensiones de «cuasidad» (*quasiness*) que ha ido en aumento a lo largo de los últimos nueve diseños es la medida en que *X* podría ser manipulada por el experimentador, es decir, en que podría introducírsela en el curso normal de los acontecimientos. Por supuesto, cuanto más ocurre así, más cerca se está de la experimentación propiamente dicha, como hemos ido viendo, en particular con referencia a los diseños 7 y 10. Los diseños 7, 10, 12, 13 (pero no 13*a*) y 14 serían aplicables tanto para *X* producidas en forma natural como para las introducidas de propósito por el experimentador. Los diseños serían más sospechosos cuando la *X* no estuviese controlada, y algunos que acaso estuviesen dispuestos a denominar cuasi*experimentos* a las versiones controladas por el experimentador, tal vez no quisieran aplicar este término a la *X* no controlada. No es nuestro ánimo hacer una cuestión por ello, pero sí destacar el valor de los análisis de datos de tipo experimental para *X* no controladas, en comparación con los ensayos evaluativos y los análisis engañosos que con demasiada frecuencia se emplean en tales situaciones. Es evidente que el diseño 15 está del todo limitado a una *X* natural, y los diseños de esta sección (aunque se los denomine de análisis de datos y no cuasiexperimentales) se hallan enclavados aún más en la situación natural. En este capítulo comenzaremos de nuevo con el análisis correlacional simple, veremos a continuación dos diseños bastante aceptables, y retornaremos por último a los experimentos *ex post facto*, que se consideran en el mejor de los casos insatisfactorios.

Correlación y causación

El diseño 3 es un diseño correlacional muy endeble, puesto que implica la comparación de solo dos unidades naturales, que difieren en la presencia y ausencia de *X*, así como tam-

bién en muchísimos otros atributos. Cada uno de ellos podría crear diferencias en las *O*, y por lo tanto cada uno ofrece una hipótesis aceptable, opuesta a la de que *X* ha producido un efecto. Nos queda una regla general: que las diferencias entre dos objetos naturales no son interpretables. Consideremos ahora esta comparación dilatada hasta el punto en que dispongamos de muchas situaciones naturales independientes de *X* y muchas otras también de no-*X*, así como diferencias concomitantes en *O*. En la medida en que las situaciones naturales de *X* varíen entre sí en sus demás atributos, esos otros atributos se tornarán menos aceptables como hipótesis rivales. Pueden establecerse, así, correlaciones de naturaleza espectacular, como las postuladas entre los fumadores empedernidos y el cáncer de pulmón. ¿Cuál es la jerarquía de esos datos como prueba de causación análoga a la suministrada por la experimentación?

Cabe ante todo hacer una reflexión positiva. Esos datos son pertinentes a las hipótesis causales en la medida en que las exponen a la refutación. Si se obtiene una correlación nula, se reduce la admisibilidad de la hipótesis. Si se produce una correlación elevada, su admisibilidad es mayor, ya que ha sobrevivido una posibilidad de refutación. Planteado el asunto en otra forma: la correlación no indica necesariamente causación, pero una ley causal del tipo que produce diferencias medias en los experimentos implica correlación. En cualquier experimento en que *X* aumente a *O*, se hallará una correlación biserial positiva entre la presencia-ausencia de *X* y los puntajes postest o los de ganancia. La ausencia de esa correlación puede eliminar muchas hipótesis causales simples y generales, relativas a los efectos principales de *X*. En este sentido, el enfoque correlacional, relativamente poco costoso, quizás ofrezca una revisión preliminar de hipótesis, y las que sobrevivan a ese proceso podrán verificarse entonces por medio de la más onerosa manipulación experimental. Katz, Maccoby y Morse [1951] han defendido esta tesis, ofreciendo una secuencia en la que los efectos del liderazgo sobre la productividad se estudiaron primero en forma correlacional, tras lo cual se verificó, por experimentación, una importante hipótesis [Morse y Reimer, 1956].

Si pasamos revista a las investigaciones sobre educación, pronto nos convenceremos de que son más los casos en que la interpretación causal de la información correlacional se exagera que aquellos en que se la desconoce, así como que suelen pasarse por alto hipótesis rivales aceptadas, y que para establecer

la antecedencia-consecuencia temporal de una relación causal es imprescindible realizar observaciones a lo largo del tiempo, cuando no apelar a la introducción experimental de *X*. Si se correlaciona, por ejemplo, el comportamiento del maestro y el alumno, nuestros estereotipos culturales casi nunca nos permitirán considerar la posibilidad de que el comportamiento del segundo provoque el del primero. Aun en una situación natural, parece hallarse implícita una prioridad temporal, y los procesos selectivos de retención pueden determinar una causalidad en sentido contrario. Consideremos, por ejemplo, posibles confirmaciones de que los inspectores que tienen a su cargo las mejores escuelas son los más cultos y que las escuelas con frecuentes cambios de inspectores tienen una moral colectiva más escasa. Es casi inevitable que extraigamos la consecuencia de que el nivel educacional de los inspectores y directivos estables *causan* mejores escuelas. La cadena causal bien podría ser a la inversa: las escuelas mejores (por el motivo que fuere) podrían ser la causa de que los hombres mejor educados permanecieran en ellas, mientras que las peores podrían inducirlos a que se sintiesen tentados a cambiar su puesto por otro mejor. De igual modo, las escuelas mejores podrían hacer que los inspectores se quedaran más tiempo en sus cargos. Aun más universal que la engañosa correlación inversa es la de una tercera variable, también conducente a error, de que los determinantes lícitos de quien está expuesto a *X* son de tal naturaleza, que producirían asimismo elevados puntajes de *O*, aun sin la presencia de *X*. Volveremos sobre estos casos en el apartado final, acerca del diseño *ex post facto*.

El experimento propiamente dicho sólo difiere de la situación correlacional en que el proceso de aleatorización destruye cualquier relación lícita entre el carácter o los antecedentes de los alumnos y su exposición a *X*. Donde se tienen pretests y no se dispone de una clara determinación de quiénes estuvieron expuestos y quiénes no, quizá sean convincentes, aun sin la aleatorización, los diseños 10 y 14. Pero para que un diseño que carece de pretest (imitando al 6) se produzca en forma natural se requieren circunstancias muy especiales, que casi nunca se dan. Así y todo, de acuerdo con nuestra tesis general relativa al aprovechamiento oportuno de las situaciones que ofrezcan datos interpretables, conviene estar alerta y con los ojos bien abiertos por si acaso se presentan. Esas situaciones serán aquellas en que parezca aceptable que la exposición a *X* no se sujete a regla alguna, sino que sea arbitraria y sin

correlación alguna con otras consideraciones. En teoría, esas decisiones de exposición arbitraria serán también muchas e independientes entre sí. Además, hay que sustentarlas por medio de cualquier otro tipo de prueba de que se disponga, por débil que sea, como en el pretest retrospectivo que analizamos más adelante. Como lo han sostenido en parte Simon [1957, págs. 10-61] y Wold [1956], la interpretación causal de una correlación simple o parcial depende tanto de la presencia de una aceptable hipótesis causal compatible como de la ausencia de hipótesis rivales lógicas para explicar la correlación sobre otros fundamentos.

Un estudio correlacional de esta índole es tan admirablemente oportuno que merece destacárselo. Barch, Trumbo y Nangle [1957] utilizaron como *X* la presencia o la ausencia de señales de giro en el automóvil que iba delante, la presencia o ausencia de las mismas señales en el auto posterior como *O*, y demostraron un significativo efecto de imitación, determinación de patrones o conformidad que concordaba con muchos estudios de laboratorio. Careciendo, como se carecía, de un pretest, la interpretación dependió del supuesto previo de que no hay relación entre las tendencias a marcar el giro en los mencionados automóviles, independientemente de la influencia ejercida por el comportamiento del automóvil que lleve la delantera. Tal como se publicó, la información parecía convincente. Nótese, sin embargo, que cualquier tercer variable que hubiera influido en forma similar sobre la frecuencia de señales de ambos pares de conductores se habría convertido en hipótesis rival aceptable. Por ende, si las condiciones atmosféricas, el grado de visibilidad, las actitudes del conductor tal como son afectadas por la hora, la presencia de un automóvil policial estacionado, etc., influyen sobre ambos conductores, y si se combinan los datos provenientes de condiciones heterogéneas en tales terceras variables, la correlación puede explicarse sin necesidad de suponer efecto alguno producido por el hecho aislado de que el auto que va delante haga la señal. Más importante como «diseño 6 natural» es el informe de Brim [1958] acerca del influjo del sexo del hermano sobre la personalidad de un niño en una familia que tiene dos hijos. La determinación del sexo puede ser una lotería casi perfecta. Hasta donde hoy se sabe, no guarda correlación alguna con los determinantes familiares, sociales o genéticos de la personalidad. La codeterminación de una tercera variable del sexo del hermano y la personalidad de un niño no es por el momento una hipótesis rival aceptable para la interpretación

causal de los interesantes descubrimientos, como tampoco lo es la causación inversa de la personalidad del niño respecto al sexo de su hermano.

El pretest retrospectivo

En muchas situaciones militares de tiempos de guerra, puede ocurrir que la asignación de hombres de igual rango y especialización a distintas unidades se haga por medio de procesos caóticos, sin consideración alguna a privilegios, preferencias o capacidades especiales. Una comparación entre las actitudes de blancos que se asignaron a unidades de infantería racialmente mixtas y las de aquellos destinados a otras integradas solo por blancos puede resultar de interés por sus determinaciones causales [Information and Education Division, 1947]. No podemos, sin duda, hacer caso omiso de estos datos, sino más bien buscar información complementaria a fin de eliminar hipótesis rivales aceptables, sin perder conciencia de las demás fuentes de invalidación. En aquel caso la entrevista «postest», no solo contenía información sobre las actitudes corrientes hacia los negros (más favorables en las compañías mixtas), sino que además requería que se recordasen las actitudes anteriores al destino actual. Aquellos «pretests retrospectivos» no arrojaron diferencia alguna entre ambos grupos, aumentando así la posibilidad de que antes de la asignación al destino no hubiera existido ninguna disparidad.

Un análisis parecido resultó importante en un estudio realizado por Deutsch y Collins [1951] comparando los ocupantes de un barrio formado por unidades integradas con los que ocupaban unidades segregadas, en momentos en que la escasez de viviendas era tal, que cabía presumir que la gente había de tomar cualquier comodidad disponible, con prescindencia casi total de sus actitudes. Teniendo tan solo mediciones postest, podría haberse considerado que las diferencias que descubrieron reflejaban sesgos de selección sobre actitudes iniciales. La interpretación de que la experiencia integrada provocó las actitudes más favorables se vio fortalecida cuando un pretest retrospectivo indicó que no había diferencias entre los dos tipos de grupos de vivienda en actitudes anteriores que se recordaran. Dados los factores autistas que, según se sabe distorsionan la memoria y los informes de las entrevistas, tales datos nunca pueden ser decisivos.

Deseamos intensamente poder trabajar con la entrevista de

pretest de entrada (y también con la asignación aleatoria a tratamientos de los moradores). Tales estudios, sin duda alguna, se están realizando. Pero hasta que se los sustituya por otros mejor fundados los descubrimientos de Deutsch y Collins, entre ellos el pretest retrospectivo, son contribuciones preciosas a una ciencia de orientación experimental en este difícil terreno.

El lector no debe pasar por alto que es probable que la memoria se incline a deformar las actitudes pasadas a fin de que concuerden con las actuales, o con lo que el morador ha llegado a considerar actitudes socialmente deseables. Parece, pues, más probable que en tales casos el sesgo de memoria se disimule, en vez de disfrazarse, como efecto significativo de *X*.

Si se continúa con los estudios comparativos de actitudes de los alumnos universitarios de primero y último año para demostrar la influencia de la institución, parece conveniente el uso de pretests retrospectivos en apoyo de las demás comparaciones como limitación parcial de las hipótesis rivales de historia, mortalidad selectiva y desvíos en la selección inicial. (Ello no quiere decir que apoyemos ninguna repetición adicional de tales estudios de corte trasversal, cuando lo que necesitamos son más estudios longitudinales, como los de Newcomb [1943], que ofrece mediciones repetidas durante el período de cuatro años, completadas en varias encuestas de corte trasversal en la forma común de una extensión a cuatro años del diseño 15. Que las tesis de doctorado, necesariamente urgidas por el tiempo, se escriban sobre otros temas.)

Estudios en panel

Las encuestas más simples recogen observaciones realizadas en un solo punto del tiempo, que a menudo ofrecen al participante la oportunidad de autoclasificarse como expuesto o no a *X*. A las correlaciones de exposición y postest que así resultan contribuye no solo el sesgo causal común (en que los determinantes de quién recibe *X* también causarían, aun sin *X*, elevados puntajes de *O*) sino también una distorsión de la memoria con respecto a *X*, dando mayor realce a la aparición espuria de causa [Stouffer, 1950, pág. 356]. Aunque estos estudios continúan apoyando las inferencias causales que justifican los presupuestos publicitarios (correlaciones entre «¿Vio usted el programa?» y «¿Compra usted el producto?»),

son pruebas muy superficiales del efecto conseguido. Introducen un nuevo factor que atenta contra la validez interna: la errónea clasificación sesgada de exposición a *X*, que no nos molestamos en incluir en nuestros cuadros.

En la metodología de la encuesta, se gana mucho con la introducción del método de panel, consistente en la repetición de entrevistas con las mismas personas. Bien practicados, los estudios en panel parecen ofrecer datos útiles para la versión más endeble del diseño 10, con *X* natural, cuando entre las dos tandas de entrevistas o cuestionarios interviene algún agente de variación, como una película cinematográfica o un contacto de asesoramiento. El estudiante de ciencias de la educación debe saber, sin embargo, que dentro de la sociología esa importante innovación metodológica suele ir acompañada por una engañosa tradición de análisis. La «tabla rotativa» [Glock, 1955], que es una tabulación cruzada con porcentajes computados con respecto a subtotales tomados como base, está muy sujeta a la confusión interpretativa de efectos de regresión con hipótesis causales, según lo señalaron Campbell y Clayton [1961]. Aun cuando se analice desde el punto de vista de las ganancias pretest-postest para un grupo expuesto frente a otro no expuesto, continúa existiendo otra fuente más sutil de sesgo. En esta modalidad de estudio en panel, la exposición a la *X* (p. ej., una película contra los prejuicios vista por mucha gente) se establece en la segunda tanda del panel en dos tandas. El diseño tiene el siguiente diagrama:

$$\begin{pmatrix} O \\ \cdots \\ O \end{pmatrix} \begin{pmatrix} X & O \\ ? & \cdots \\ & O \end{pmatrix}$$

Diseño en panel con dos tandas (inaceptable)

Aquí, los paréntesis indican la ocurrencia de *O* o *X* en la misma entrevista; el signo de interrogación, ambigüedad de clasificación en grupos *X* y no-*X*. A diferencia del diseño 10, la *X* está correlacionada con las *O* del pretest (en que los de menos prejuicios realizan los mayores esfuerzos por ir a ver la película). Pero, además, aunque *X* no hubiera tenido ningún efecto real sobre *O*, la correlación entre *X* y los postests sería mayor que entre *X* y el pretest solo, porque se producen en la misma entrevista.

En la investigación con pruebas y mediciones es bastante habitual que se observe una mayor tendencia a la correlación entre dos puntos cualesquiera incluidos en el mismo cuestionario que si se encontraran en distintos cuestionarios. Stockford y Bissell [1949] comprobaron que los ítems adyacentes se correlacionaban más que los no adyacentes, incluso en el mismo instrumento. Las pruebas administradas en el mismo día tienen mayor correlación que aquellas que se aplican en días distintos. En el estudio en panel que comentamos [Glock, 1955], ambas entrevistas se produjeron con unos 8 meses de intervalo. Las fuentes de correlación que destacan las que aparecen en una misma entrevista y oscurecen las existentes en entrevistas separadas no solo incluyen fluctuaciones autónomas en los prejuicios, sino también diferencias en los entrevistadores. Los inevitables errores cometidos por el entrevistador, así como las inexactas manifestaciones del entrevistado al reidentificar a participantes anteriores, provocan que algunos de los pares pretest-postest deriven, en realidad, de personas distintas. La más elevada correlación resultante *X*-postest implica que habrá una menor regresión del informe de *X* al postest que al pretest y, por tanto, que las diferencias postest en *O* serán mayores que las pretest. Esto se traducirá (si no se ha producido ningún incremento de población) en una seudoganancia para los autoclasificados como expuestos y una seudopérdida para los que se clasificaron como no expuestos. Este resultado se confundirá por lo común con una confirmación de la hipótesis de que *X* ha tenido un efecto [véase Campbell y Clayton, 1961, para los detalles de esta argumentación].

Para evitar esta fuente espuria de mayor correlación, se podría determinar la exposición a *X* en forma independiente de la entrevista, o en una tanda intermedia de entrevistas separadas. En este último caso, aunque se conservase un recuerdo sesgado de exposición, ello no produciría artificialmente ninguna correlación *X*-postest más elevada que la *X*-pretest. Un diseño de esta índole adoptaría la siguiente forma:

$$\begin{matrix} O \\ \cdots \\ O \end{matrix} \quad \begin{matrix} X \\ \cdot ? \cdot \cdot \end{matrix} \quad \begin{matrix} O \\ \cdots \\ O \end{matrix}$$

El cuadro de dieciséis partes de Lazarsfeld

Otra ingeniosa aplicación cuasiexperimental de la información de panel, introducida por Lazarsfeld alrededor de 1948 en un informe mimeografiado titulado «The mutual effect of statistical variables» (El efecto mutuo de las variables estadísticas), tuvo por objetivo en un primer momento la obtención de un índice del sentido (y fuerza) de la causación existente entre dos variables. Ese análisis se designa en la actualidad con el nombre de «Cuadro de dieciséis partes» [p. ej., Lipset, Lazarsfeld, Barton y Linz, 1954, págs. 1160-63], y se emplea por lo común para averiguar la fuerza o profundidad relativa de varias actitudes, más que para inferir el «sentido de causación». Este último propósito es el que lo convierte en cuasiexperimental.

Supongamos que en determinada ocasión podemos clasificar el comportamiento de cien maestros como «cálido» o «frío», y el correspondiente a sus alumnos como de «interesados» o «no interesados».

Al hacerlo así, descubrimos una correlación positiva: los maestros cálidos tienen clases interesadas. Cabe plantearse ahora el interrogante de si es la calidez del maestro la que provoca el interés de la clase, o viceversa. Aunque nuestras expectativas culturales nos predisponen en favor de la primera interpretación, puede presentarse también un argumento nada desdeñable en favor de la segunda. (Interviene, sin duda, un efecto de causación recíproca.) Un estudio en panel agregaría datos pertinentes, al volver a ponderar las mismas variables en una segunda sesión, con los mismos maestros y cursos. (Dos niveles de medición para dos variables generan cuatro tipos de reacciones para cada sesión, o sea 4×4 posibles configuraciones de reacción para ambas acciones, produciendo el cuadro de dieciséis partes).

Con fines simplemente ilustrativos, supongamos ahora el siguiente resultado:

Primera sesión.

Alumnos	*Maestros*	
	Frío	Cálido
Interesados	20	30
No interesados	30	20

Segunda sesión.

Alumnos	*Maestros* Frío	Cálido
Interesados	10	40
No interesados	40	10

Saltan a la vista tanto la posibilidad de error de la información correlacional ordinaria como el ingenio del análisis de Lazarsfeld, si notamos que entre los desplazamientos que habrían posibilitado la trasformación se dan los siguientes opuestos polares:

La calidez del maestro provoca interés en los alumnos.

Alumnos	*Maestros* Frío	Cálido
Interesados	10 (10)	30
No interesados	30	(10) 10

El interés de los alumnos provoca calidez en el maestro.

Alumnos	*Maestros* Frío	Cálido
Interesados	10 (10) →	30
No interesados	30 ←	(10) 10

Hemos considerado aquí solo los cambios que aumentan la intercorrelación, soslayando las inevitables oscilaciones. Así, en este diagrama, a diferencia del de Lazarsfeld, no presentamos más que 8 de los 16 casilleros de su cuadro en dieciséis partes, limitándonos a los cuatro tipos estables (repetidos tanto en el diagrama superior como en el inferior) y los cuatro tipos de desplazamientos que aumentarían la correlación (dos arriba y dos abajo). Los cuatro tipos de desplazamientos podrían, por supuesto, producirse a la vez, y cualquier inferencia a propósito del sentido de la causación se fundaría en una

preponderancia del uno sobre el otro. Estos diagramas representan los dos resultados más claros posibles. De producirse uno de ellos, el examen de los sujetos que se desplazan, posibilitado por la recopilación de datos tipo panel (imposible si en cada caso actuasen distintos alumnos y maestros), parece otorgar gran admisibilidad a una inferencia causal monodireccional. Para los que se desplazaron, pueden notarse la dimensión temporal y el sentido del cambio. De verificarse el caso indicado en primer término, sería poco probable que los alumnos estuvieran cambiando de maestros, y muy probable que los maestros estuvieran cambiando de alumnos, al menos en esos veinte cursos cambiantes.

Aunque los sociólogos dejan el análisis al nivel dicotómico, estos requisitos pueden formularse de nuevo en forma más general, como correlaciones desfasadas en el tiempo, donde el «efecto» debería tener una correlación más elevada con una «causa anterior» que con una «ulterior»; es decir, $rx_1\, o_2 > rx_2\, o_1$. Tomando el caso en que los maestros son los causantes de la conducta de los alumnos, obtenemos:

Alumnos segunda vez	*Maestros primera vez* Frío	Cálido
Interesados	10	40
No interesados	40	10

Alumnos primera vez	*Maestros segunda vez* Frío	Cálido
Interesados	20	30
No interesados	30	20

En este caso el ejemplo parece una reformulación trivial de los cuadros originales, ya que los maestros no cambiaron en absoluto. Sin embargo, es tal vez la mejor forma general de análisis. Nótese que, pese a ser aceptable, tal vez no debería utilizarse el argumento $rx_1\, o_2 > rx_1\, o_1$, a causa de las muchas fuentes no pertinentes de correlación que se producen entre conjuntos de datos tomados en la misma sesión, que inflarían el valor $rx_1\, o_1$. Téngase en cuenta que el $rx_1\, o_2 > rx_2\, o_1$ sugerido no otorga a ninguna de las correlaciones la menor ventaja a este respecto.

¿Cuáles son los inconvenientes de este diseño? La aplicación

de tests, porque su repetición puede traducirse de manera bastante general en correlaciones más elevadas entre las variables correlacionadas. El $rx_1\ o_1 < rx_2\ o_2$ preliminar puede explicarse sobre esta base. No obstante, ello no explicaría con facilidad el hallazgo de $rx_1\ o_2 > rx_2\ o_1$, a menos que fuese aceptable un efecto de interacción o aplicación de tests peculiar de solo una de las variables.

La regresión parece constituir un problema menor para este diseño que para el estudio en panel con dos tandas rechazado antes, porque tanto *X* como *O* se evalúan en ambas tandas, y por consiguiente la clasificación en tales términos resulta simétrica. Sin embargo, para el análisis dicotómico tipo Lazarsfeld, la regresión pasa a ser un problema si los marginales de cualquiera de las variables presentan una asimetría grave (p. ej., divisiones 10-90 en vez de las 50-50 utilizadas en estos ejemplos). El análisis de correlaciones entre variables continuas, empleando todos los casos, no parecería estar en conflicto con los mecanismos de regresión. La maduración diferencial en ambas variables, o los efectos diferenciales de la historia, podrían ser efectos de interacción que pusieran en peligro la validez interna. En cuanto a la externa, son de aplicación las precauciones habituales, con particular insistencia en la interacción selección-*X* en el sentido de que el efecto se ha observado solo a propósito de la subpoblación que se desplaza.

Si bien en la mayor parte de las situaciones de enseñanza se dispondría de los diseños 10 o 14 para el tipo de problema planteado en nuestro ejemplo (y serían preferibles) es probable que existan situaciones en las cuales debería considerarse este análisis. El doctor Winfred F. Hill, por ejemplo, ha recomendado su aplicación a los datos obtenidos sobre el comportamiento de padres e hijos en estudios longitudinales.[1]

Cuando se generaliza a datos no dicotómicos, el nombre «Cuadro en dieciséis partes» deja de ser apropiado; recomendamos que se lo denomine «Correlación en panel con desfasaje cruzado».

Análisis «ex post facto»

En la actualidad, la frase «experimento *ex post facto*» designa los esfuerzos para simular la experimentación por medio de

1 Comunicación personal.

un proceso en el que se intenta una situación de diseño 3 con miras a lograr una ecuación pre-*X*, empleando un proceso de equiparación en atributos pre-*X*. El modo de análisis y su nombre los introdujo por primera vez Chapin [Chapin y Queen, 1937]. Más adelante han expuesto con amplitud este diseño Greenwood [1945] y Chapin [1947, 1955]. Aunque estas referencias provienen de la sociología y no de la pedagogía, y consideramos que el análisis conduce a error, entendemos que corresponde exponerlo también en esta obra. Constituye uno de los esfuerzos más amplios con miras al diseño cuasiexperimental. Los ejemplos proceden con frecuencia del ámbito educacional. La lógica utilizada y los errores en que se incurre, son también frecuentes en la investigación pedagógica.

En un típico estudio *ex post facto* [Chapin, 1955, págs. 99-124], la *X* era la educación recibida en la escuela secundaria (sobre todo en sus últimos años) y las *O* se relacionaban con el éxito y el ajuste comunitario diez años después, juzgados sobre la base de datos obtenidos en entrevistas personales. La equiparación se hizo en aquella oportunidad recurriendo a los archivos escolares (aunque en estudios análogos, más endebles todavía, aquellos hechos pre-*X* se obtenían en las entrevistas post-*X*). En principio los datos indicaron que quienes completaban la escuela secundaria habían tenido más éxito, pero también había influido en ello el mejor puntaje obtenido en la escuela primaria, la ocupación de los padres en niveles superiores, la menor edad, los mejores vecindarios, etc. Esos antecedentes, pues, podrían haber sido la causa, tanto de la finalización de la escuela secundaria como del éxito posterior.

¿Ejerció la escuela algún influjo adicional por encima del mejor comienzo ofrecido por esos factores ambientales? La «solución» de Chapin a este interrogante fue examinar subconjuntos de estudiantes equiparados en todos aquellos factores ambientales, pero con diferencias al concluir la escuela secundaria. El agregado de cada factor de equiparación redujo a su vez la discrepancia postest entre los grupos *X* y no *X*, pero una vez realizadas todas las equiparaciones quedó una diferencia significativa. Chapin llegó a la conclusión, si bien cauta, de que la educación había tenido un efecto. Un universo inicial de 2.127 estudiantes se redujo a 1.194 entrevistas completadas sobre casos con antecedentes adecuados. El ajuste redujo los casos utilizables a 46, es decir, 23 graduados y 23 no graduados, menos del 4 % de los entrevistados.

Chapin sostiene correctamente que 46 casos comparables son

preferibles a 1.194 no comparables, sobre fundamentos similares a nuestro énfasis relativo a la prioridad de la validez interna sobre la externa. Lo lamentable es que sus 46 casos tampoco son comparables, y lo que es más grave todavía: aun admitiendo su defectuosa argumentación, la reducción era innecesaria.

Incurrió en una grave *sub*equiparación por dos razones distintas. Su primera fuente de subequiparación fue que la equiparación está sujeta a regresión diferencial, la que en este caso produciría por cierto una diferencia final en el sentido obtenido (de la manera indicada por R. L. Thorndike, 1942 y analizada a propósito de la equiparación en el diseño 10). El sentido del seudoefecto de la regresión relativa a medias grupales después de la equiparación es en este caso seguro, pues las diferencias en los factores de equiparación entre los que lograron éxito frente a los que no lo tuvieron tienen el mismo sentido para cada factor que las diferencias entre los que completaron la escuela secundaria y los que la abandonaron antes de finalizar sus estudios.

Cada determinante de exposición a X es, de manera similar y aun sin X, un determinante de O. Todas las variables equiparadas correlacionan con X y O en el mismo sentido. Aunque bien podría no ocurrir así en todas las variables de todos los estudios *ex post facto*, sí acaece, si no en todos, en la mayor parte de los ejemplos publicados. Este error y la reducción en el número de casos pueden evitarse por medio de la estadística moderna, que elude el error de equiparación en el diseño 10.

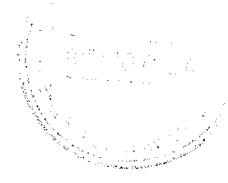
Las variables de equiparación podrían ser utilizadas en su totalidad como covariables en un análisis de covariancia con covariables múltiples. Estimamos con toda seriedad que ese análisis eliminaría los efectos aparentemente significativos en los estudios específicos presentados por Chapin. (Véase, sin embargo, Lord [1960], por su crítica del análisis de covariancia para problemas de esta índole.) Pero hay otra inevitable fuente de subequiparación en la configuración de Chapin. Greenwood [1945] la designa con el nombre de autoselección de exposición o no exposición. La exposición es consecuencia lógica de muchos antecedentes. En el caso del abandono de la escuela secundaria antes de finalizarla, sabemos que son innumerables los determinantes posibles, además de aquellos sobre los cuales se hizo la equiparación. Podemos suponer, con gran seguridad, que casi todos ellos tendrán un efecto similar sobre éxitos ulteriores, independientemente de su efecto por medio de X. Este solo hecho asegura que la subequiparación sobrepasará el efecto de regresión por equiparación. Aun con el predictor pre-X y el análisis de covariancia de O, solo es interpretable un efecto significativo de tratamiento cuando se han incluido *todas* las variables equiparadas que contribuyen en forma conjunta.

7. Comentarios finales

Esta obra es ya en sí una exposición condensada. Un resumen de ella puede inducir a engaño. En este sentido, parece imprescindible una advertencia final a propósito de la tendencia a utilizar con ese fin los cuadros 1, 2 y 3, de aparente pero falsa conveniencia. Esos cuadros, en calidad de bosquejo recurrente, han contribuido en parte a ordenar la obra haciéndola menos reiterativa. Pero la colocación de signos positivos, negativos e interrogativos ha sido siempre equívoca y, por lo común, constituye un mal resumen del correspondiente análisis. Es probable que en cualquier ejecución particular de un diseño, la fila de comprobación resulte distinta de la que aparece en el correspondiente cuadro.

Por ejemplo, la solución de empate del diseño 6 a la cual aludimos al pasar en el diseño cuasiexperimental 16 tiene, según lo allí expuesto, dos signos interrogativos y uno negativo que no aparecen en el cuadro 1 en la fila del diseño 6. La mejor forma de utilizar los cuadros es hacerlo a manera de otros tantos bosquejos de un cuidadoso estudio de los detalles particulares de un experimento durante la etapa de su planificación. Del mismo modo, esta obra no pretende sustituir con el dogma de *los 13 diseños aceptables* otro dogma anterior del diseño o par de diseños aceptables. Más bien habría que estimular una actitud abierta hacia la indagación de los nuevos mecanismos de obtención de datos, y un nuevo análisis acerca de algunas de las imperfecciones que acompañan a la aplicación rutinaria de los tradicionales.

Por último, hemos visto en este trabajo distintas alternativas sobre los dispositivos o diseños experimentales, con particular referencia a los problemas de control de variables externas y amenazas a la validez. Hay que distinguir entre validez interna y externa, o posibilidad de generalizar. Se han empleado ocho clases de amenazas a la validez interna y cuatro a la externa, para evaluar dieciséis diseños experimentales y unas cuantas variaciones sobre ellos. Tres de esos diseños se han clasificado como preexperimentales y se los ha empleado sobre todo para ilustrar los factores de validez que requieren control. Tres de ellos se consideraron diseños experimentales propiamente dichos. Y de diez se ha dicho que son cuasiexperimentos pues carecen de un control perfecto, si bien merecen que se los adopte cuando no haya nada mejor. Para interpretar los resultados de esos experimentos, es de particular importancia la lista de verificación de factores de validez. En general, hemos llamado la atención sobre la posibilidad de utilizar creativamente las características peculiares de cualquier situación concreta de investigación al diseñar pruebas específicas de las hipótesis causales.



Referencias bibliográficas

Allport, F. H.
1920. «The influence of the group upon association and thought», *J. Exp. Psychol.*, vol. 3, págs. 159-82.

Anastasi, A.
1958. *Differential psychology*, Nueva York: Macmillan, 3a. ed. (*Psicología diferencial*, Madrid: Aguilar, 1964.)

Anderson, N. H.
1959. «Test of a model for opinion change», *J. Abnorm. Soc. Psychol.*, vol. 59, págs. 371-81.

Barch, A. M., Trumbo, D. y Nangle, J.
1957. «Social setting and conformity to a legal requirement», *J. Abnorm. Soc. Psychol.*, vol. 55, págs. 396-98.

Blalock, H. M.
1964. *Causal inferences in nonexperimental research*, Chapel Hill: University of North Carolina Press.

Boring, E. G.
1954. «The nature and the history of experimental control», *Amer. J. Psychol.*, vol. 67, págs. 573-89.

Box, G. E. P.
1967. «Bayesian approaches to some bothersome problems in data analysis», en J. C. Stanley, ed., *Improving experimental design and statistical analysis*, Chicago: Rand McNally.

Box, G. E. P. y Tiao, G. C.
1965. «A change in level of a non-stationary time series», *Biometrika*, vol. 52, págs. 181-92.

Brim, O. G.
1958. «Family structure and sex role learning by children: A further analysis of Helen Koch's data», *Sociometry*, vol. 21, págs. 1-16.

Brolyer, C. R., Thorndike, E. L. y Woodyard, E.
1927. «A second study of mental discipline in high school studies», *J. Educ. Psychol.*, vol. 18, págs. 377-404.

Brownlee, K. A.
1960. *Statistical theory and methodology in science and engineering*, Nueva York: Wiley.

Brunswik, E.
1956. *Perception and the representative design of psychological experiments*, Berkeley: University of California Press, 2a. ed.

Campbell, D. T.
1957. «Factors relevant to the validity of experiments in social settings», *Psychol. Bull.*, vol. 54, págs. 297-312.
1959. «Methodological suggestions from a comparative psychology of knowledge processes», *Inquiry*, vol. 2, págs. 152-82.
1960. «Recommendations for APA test standards regarding construct, trait, or discriminant validity», *Amer. Psychologist* vol. 15, págs. 546-53.
1963. «From description to experimentation: Interpreting trends as cuasi experiments», en C. W. Harris, ed., *Problems in measuring change*, Madison: University of Wisconsin Press, págs. 212-42.
1967. «Administrative experimentation, institutional records, and nonreactive measures», en J. C. Stanley, ed., *Improving experimental design and statistical analysis*, Chicago: Rand McNally.
«Quasi-experimental designs for use in natural social settings», en D. T. Campbell, *Experimenting, validating, knowing: Problems of method in the social sciences*, Nueva York: McGraw-Hill, en preparación.

Campbell, D. T. y Clayton, K. N.
1961. «Avoiding regression effects in panel studies of communication impact», *Stud. Pub. Commun.*, nº 3, págs. 99-118.

Campbell, D. T. y Fiske, D. W.
1959. «Convergent and discriminant validation by the multitrait-multimethod matrix», *Psychol. Bull.*, vol. 56, págs. 81-105.

Campbell, D. T. y McCormack, T. H.
1957. «Military experience and attitudes toward authority», *Amer. J. Sociol.*, vol. 62, págs. 482-90.

Cane, V. R. y Heim, A. W.
1950. «The effects of repeated testing: III. Further experi

ments and general conclusions», *Quart. J. Exp. Psychol.*, vol. 2, págs. 182-95.

Cantor, G. N.
1956. «A note on a methodological error commonly committed in medical and psychological research», *Amer. J. Ment. Defic.*, vol. 61, págs. 17-18.

Cochran, W. G. y Cox, G. M.
1957. *Experimental designs*, Nueva York: Wiley, 2a. ed.

Collier, R. M.
1944. «The effect of propaganda upon attitude following a critical examination of the propaganda itself», *J. Soc. Psychol.*, vol. 20, págs. 3-17.

Collier, R. O., h.
1960. «Three types of randomization in a two-factor experiment», Minneapolis: edición del autor, 1960 (fotocopia).

Cornfield, J. y Tukey, J. W.
1956. «Average values of mean squares in factorials», *Ann. Math. Statist.*, vol. 27, págs. 907-49.

Cox, D. R.
1951. «Some systematic experimental designs», *Biometrika*, vol. 38, págs. 312-23.
1957. «The use of a concomitant variable in selecting an experimental design», *Biometrika*, vol. 44, págs. 150-58.
1958. *Planning of experiments*, Nueva York: Wiley.

Crook, M. N.
1937. «The constancy of neuroticism scores and self-judgments of constancy», *J. Psychol.*, vol. 4, págs. 27-34.

Chapin, F. S.
1947. *Experimental designs in sociological research*, Nueva York: Harper, 1947. ed. rev., 1955.

Chapin, F. S. y Queen, S. A.
1937. *Research memorandum on social work in the depression*, Nueva York: boletín nº 39, del Social Science Research Council.

Chernoff, H. y Moses, L. E.
1959. *Elementary decision theory*, Nueva York: Wiley. (*Teoría y cálculo elemental de las decisiones*, México: CECSA, 1967.)

Deutsch, M. y Collins, M. E.
1951. *Interracial housing: A psychological evaluation of a social experiment*, Minneapolis: University of Minnesota Press.

Duncan, C. P., O'Brien, R. B., Murray, D. C., Davis, L. y Gilliland, A. R.
1957. «Some information about a test of psychological misconceptions», *J. Gen. Psychol.*, vol. 56, págs. 257-60.

Ebbinghaus, H.
1913. *Memory*, Nueva York: Columbia University, Teachers College.

Edwards, A. L.
1960. *Experimental design in pyschological research*, Nueva York, Rinehart, ed. rev.

Farmer, E., Brooks, R. C. y Chambers, E. G.
1923. *A comparison of different shift systems in the glass trade*. Rep. 24, Medical Research Council, Industrial Fatigue Research Board, Londres: His Majesty's Stationery Office.

Feldt, L. S.
1958. «A comparison of the precision of three experimental designs employing a concomitant variable», *Psychometrika*, vol. 23, págs. 335-53.

Ferguson, G. A.
1959. *Statistical analysis in psychology and education*, Nueva York: McGraw-Hill.

Fisher, R. A.
1925. *Statistical methods for research workers*, Londres: Oliver & Boyd, 1a. ed. (*Tablas estadísticas para investigadores científicos*, Madrid: Aguilar, 2a. ed., 1954.)
1926. «The arrangement of field experiments», *J. Min. Agriculture*, vol. 33, págs. 503-13; reimpreso en *Contributions to mathematical statistics*, Nueva York: Wiley, 1950.
1935. *The design of experiments*, Londres: Oliver & Boyd, 1a. ed.

Glass, G. V.
1965. «Evaluating testing, maturation, and treatment effects in a pretest posttest quasi-experimental design», *Amer. Educ. Res. J.*, vol. 2, págs., 83-7.

Glickman, S. E.
1961. «Perseverative neural processes and consolidation of the memory trace», *Psychol. Bull.*, vol. 58, págs. 218-33.

Glock, C. Y.
1955. «Some applications of the panel method to the study

of social change», en P. F. Lazarsfeld y M. Rosenberg, eds., *The language of social research*, Glencoe Ill.: Free Press, págs. 242-49.
1958. «The effects of re-interviewing in panel research», copia litográfica de un capítulo de P. F. Lazarsfeld, ed., *The study of short run social change*, en preparación.

Good, C. V. y Scates, D. E.
1954. *Methods of research*, Nueva York: Appleton-Century-Crofts.

Grant, D. A.
1956. «Analysis-of-variance tests in the analysis and comparison of curves», *Psychol. Bull.*, vol. 53, págs. 141-54.

Green, B. F. y Tukey, J. W.
1960. «Complex analyses of variance: General problems», *Psychometrika*, vol. 25, págs. 127-52.

Greenwood, E.
1945. *Experimental sociology: A study in method*, Nueva York: King's Crown Press.

Guetzkow, H., Kelly, E. L. y McKeachie, W. J.
1954. «An experimental comparison of recitation, discussion, and tutorial methods in college teaching», *J. Educ. Psychol.*, vol. 45, págs. 193-207.

Hammond, K. R.
1954. «Representative vs. systematic design in clinical psychology», *Psychol. Bull.*, vol. 51, págs. 150-59.

Hanson, N. R.
1958. *Patterns of discovery*, Cambridge, Inglaterra: University Press.

Hovland, C. I., Janis, I. L. y Kelley, H. H.
1953. *Communication and persuasion*, New Haven, Conn.: Yale University Press.

Hovland, C. I., Lumsdaine, A. A. y Sheffield, F. D.
1949. *Experiments on mass communication*, Princeton N. J.: Princeton University Press.

Johnson, P. O.
1949. *Statistical methods in research*, Nueva York: Prentice-Hall.

Johnson, P. O. y Jackson, R. W. B.
1959. *Modern statistical methods: Descriptive and inductive*, Chicago: Rand McNally.

Jost, A.
1897. «Die Assoziationsfestigkeit in ihrer Abhängigkeit von der Verteilung der Wiederholungen», *Z. Psychol. Physiol. Sinnesorgane*, vol. 14, págs. 436-72.

Kaiser, H. F.
1960. «Directional statistical decisions», *Psychol. Rev.*, vol. 67, págs. 160-67.

Katz, D., Maccoby, N. y Morse, N. C.
1951. *Productivity, supervision, and morale in an office situation*, Ann Arbor: University of Michigan, Survey Research Center.

Kempthorne, O.
1952. *The desing and analysis of experiments*, Nueva York: Wiley.
1955. «The randomization theory of statistical inference», *J. Amer. Statist. Ass.*, vol. 50, págs. 946-67; 1956, vol. 51, pág. 651.
1961. «The design and analysis of experiments, with some reference to educational research», en R. O. Collier y S. M. Elam, eds., *Research design and analysis: The second annual Phi Delta Kappa symposium on educational research*, Bloomington, Ind.: Phi Delta Kappa, págs. 97-133.

Kendall, M. G. y Buckland, W. R.
1957. *A dictionary of statistical terms*, Londres: Oliver & Boyd.

Kennedy, J. L. y Uphoff, H. F.
1939. «Experiments on the nature of extra-sensory perception. III. The recording error criticisms of extrachance scores», *J. Parapsychol.*, vol. 3, págs. 226-45.

Kerr, W. A.
1945. «Experiments on the effect of music on factory production», *Appl. Psychol. Monogr.*, nº 5.

Lana, R. E.
1959*a* «Pretest-treatment interaction effects in attitudinal studies», *Psychol. Bull.*, vol. 56, págs. 293-300.
1959*b*. «A further investigation of the pretest-treatment interaction effect, *J. Appl. Psychol.*, vol. 43, págs. 421-22.

Lana, R. E. y King, D. J.
1960. Learning factors as determiners of pretest sensitization», *J. Appl. Psychol.*, vol. 44, págs. 189-91.

Lindquist, E. F.
1940. *Statistical analysis in educational research*, Boston: Houghton Mifflin.
1953. *Design and analysis of experiments in psychology and education*, Boston: Houghton Mifflin.

Lipset, S. M., Lazarsfeld, P. F., Barton, A. H. y Linz, J.
1954. «The psychology of voting: An analysis of political behavior», en G. Lindzey, ed., *Handbook of social psychology*, Cambridge, Mass.: Addison-Wesley, págs. 1124-75. (*Manual de psicología social*, Buenos Aires: Paidós, en preparación.)

Lord, F. M.
1956. «The measurement of growth», *Educ. Psychol. Measmt.*, vol. 16, págs. 421-37.
1958. «Further problems in the measurement of growth», *Educ. Psychol. Measmt.*, vol. 18, págs. 437-51.
1960. «Large-sample covariance analysis when the control variable is fallible», *J. Amer. Statist. Ass.*, vol. 55, págs. 307-21.

Lubin, A.
1961. «The interpretation of significant interaction», *Educ. Psychol. Measmt.*, vol. 21, págs. 807-17.

Maxwell, A. E.
1958. *Experimental design in psychology and the medical sciences*, Londres: Methuen.

McCall, W. A.
1923. *How to experiment in education*, Nueva York: Macmillan.

McNemar, Q.
1940. «A critical examination of the University of Iowa studies of environmental influences upon the I. Q.», *Psychol. Bull.*, vol. 37, págs. 63-92.
1958. «On growth measurement», *Educ. Psychol. Measmt.*, vol. 18, págs. 47-55.
1962. *Psychological statistics*, Nueva York: Wiley, 3a. ed.

Meehl, P. E.
1954. *Clinical versus statistical prediction*, Minneapolis: University of Minnesota Press.

Monroe, W. S.
1938. «General methods: Classroom experimentation», en G. M. Whipple, ed., *Yearb. Nat. Soc. Stud. Educ.*, vol. 37, part. II, págs. 319-27.

Mood, A. F.
1950. *Introduction to the theory of statistics*, Nueva York: McGraw-Hill (*Introducción a la teoría de la estadística*, Madrid: Aguilar, 2a. ed., 1969.)

Moore, H. T.
1921. «The comparative influence of majority and expert opinion», *Amer. J. Psychol.*, vol. 32, págs. 16-20.

Morse, N. C. y Reimer, E.
1956. «The experimental change of a major organizational variable», *J. Abnorm. Soc. Psychol.*, vol. 52, págs. 120-29.

Myers, J. L.
1959. «On the interaction of two scaled variables», *Psychol. Bull.*, vol. 56, págs. 384-91.

Newcomb, T. M.
1943. *Personality and social change*, Nueva York: Dryden.

Neyman, J.
1960. «Indeterminism in science and new demands on statisticians», *J. Amer. Statist. Ass.*, vol. 55, págs. 625-39.

Nunnally, J.
1960. «The place of statistics in psychology», *Educ. Psychol. Measmt.*, vol. 20, págs. 641-50.

Page, E. B.
1958. «Teacher comments and student performance: A seventy-four classroom experiment in school motivation», *J. Educ. Psychol.*, vol. 49, págs. 173-81.

Pearson, H. C.
1912. «Experimental studies in the teaching of spelling», *Teachers Coll. Rec.*, vol. 13, págs. 37-66.

Pelz, D. C. y Andrews, F. M.
1964. «Detecting causal priorities in panel study data», *Amer. Sociol. Rev.*, vol. 29, págs. 836-48.

Peters, C. C. y Van Voorhis, W. R.
1940. *Statistical procedures and their mathematical bases*, Nueva York: McGraw-Hill.

Piers, E. V.
1954. «Effects of instruction on teacher attitudes: Extended control-group design», tesis inédita de doctorado, George Peabody Coll.
1955. «Abstract», *Bull. Maritime Psychol. Ass.*, págs. 53-56.

Popper, K. R.
1959. *The logic of scientific discovery*, Nueva York: Basic Books. (*La lógica de la investigación científica*, Madrid: Tecnos, 1965.)

Rankin, R. E. y Campbell, D. T.
1955. «Galvanic skin response to negro and white experimenters», *J. Abnorm. Soc. Psychol.*, vol. 51, págs. 30-3.

Reed, J. C.
1956. «Some effects of short term training in reading under conditions of controlled motivation», *J. Educ. Psychol.*, vol. 47, págs. 257-64.

Rogers, C. R. y Dymond, R. F.
1954. *Psychotherapy and personality change*, Chicago: University of Chicago Press.

Rosenthal, R.
1959. «Research on experimenter bias», trabajo leído en la American Psychological Association, Cincinnati.

Roy, S. N. y Gnanadesikan, R.
1959. «Some contributions to ANOVA in one or more dimensions: I and II», *Ann. Math. Statist.*, vol. 30, págs. 304-17, 318-40.

Rozeboom, W. W.
1960. «The fallacy of the nullhypothesis significance test» *Psychol. Bull.*, vol. 57, págs. 416-28.

Rulon, P. J.
1941. «Problems of regression», *Harvard Educ. Rev.*, vol. 11, págs. 213-23.

Sanford, F. H. y Hemphill, J. K.
1952. An evaluation of a brief course in psychology at the U. S. Naval Academy», *Educ. Psychol. Measmt.*, vol. 12, págs. 194-216.

Scheffé, H.
1956. «Alternative models for the analysis of variance», *Ann. Math. Statist.*, vol. 27, págs. 251-71.

Selltiz, C., Jahoda, M., Deutsch, M. y Cook, S. W.
1959. *Research methods in social relations*, Nueva York: Holt-Dryden, ed. rev. (*Métodos de investigación en las relaciones sociales*, Madrid: Rialp, 2a. ed., 1965.)

Siegel, A. E. y Siegel, S.

1957. «Reference groups, membership groups, and attitude change», *J. Abnorm. Soc. Psychol.*, vol. 55, págs. 360-64.

Simon, H. A.
1957. *Models of man*, Nueva York: Wiley.

Smith, H. L. y Hyman, H.
1950. «The biasing effect of interviewer expectations on survey results», *Publ. Opin. Quart.*, vol. 14, págs. 491-506.

Sobol, M. G.
1959. «Panel mortality and panel bias», *J. Amer. Statist. Ass.*, vol. 54, págs. 52-68.

Solomon, R. L.
1949. «An extension of control group design», *Psychol. Bull.*, vol. 46, págs. 137-50.

Sorokin, P. A.
1930. «An experimental study of efficiency of work under various specified conditions», *Amer. J. Sociol.*, vol. 35, págs. 765-82.

Stanley, J. C.
1955. «Statistical analysis of scores from counterbalanced tests», *J. Exp. Educ.*, vol. 23, págs. 187-207.
1956. «Fixed random, and mixed models in the analysis of variance as special cases of finite model III», *Psychol. Rep.*, vol. 2, pág. 369.
1957*a*. «Controlled experimentation in the classroom», *J. Exp. Educ.*, vol. 25, págs. 195-201.
1957*b*. «Research methods: Experimental design», *Rev. Educ. Res.*, vol. 27, págs. 449-59.
1960. «Interactions of organisms with experimental variables as a key to the integration of organismic and variable-manipulating research», en E. M. Huddleston, ed., *Yearb. Nat. Counc. Measmt. used in Educ.*, págs. 7-13.
1961*a*. «Analysis of a double nested design», *Educ. Psychol. Measmt.*, vol. 21, págs. 831-37.
1961*b*. «Studying status vs. manipulating variables», en R. O. Collier y S. M. Elam, eds., *Research design and analysis: The second Phi Delta Kappa symposium on educational research*. Bloomington, Ind.: Phi Delta Kappa, págs. 173-208.
1961*c*. «Analysis of unreplicated threeway classifications, with applications to rater bias and trait independence», *Psychometrika*, vol. 26, págs. 205-20.
1962. «Analysis-of-variance principles applied to the grading of essay tests», *J. Exp. Educ.*, vol. 30, págs. 279-83.

1965. «Quasi-experimentation», *Sch. Rev.*, vol. 73, págs. 197-205.
1966*a*. «A common class of pseudo-experiments», *Amer. Educ. Res. J.*, vol. 3, págs. 79-87.
1966*b*. «The influence of Fischer's *The design of experiments* on educational research thirty years later», *Amer. Educ. Res. J.*, vol. 3, págs. 223-29.
1966*c*. «Rice as a pioneer educational researcher», *J. Educ. Measmt.*, vol. 3, págs. 135-39.

Stanley, J. C. y Beeman, E. Y.
1956. «Interaction of major field of study with kind of test», *Psychol. Rep.*, vol. 2, págs. 333-36.
1958. «Restricted generalization, bias, and loss of power that may result from matching groups», *Psychol. Newsltr.*, vol. 9, págs. 88-102.

Stanley, J. C. y Wiley, D. E.
1962. *Development and analysis of experimental designs for ratings*, Madison, Wis.: edición de los autores.

Stanton, F. y Baker, K. H.
1942. «Interviewer-bias and the recall of incompletely learned materials», *Sociometry*, vol. 5, págs. 123-34.

Star, S. A. y Hughes, H. M.
1950. «Report on an educational campaign: The Cincinnati plan for the United Nations», *Amer. J. Sociol.*, vol. 55, págs. 389-400.

Stockford, L. y Bissell, H. W.
1949. «Factor involved in establishing a merit-rating scale», *Personnel*, vol. 26, págs. 94-116.

Stouffer, S. A., ed.
1949. *The American soldier*, Princeton, N. J.: Princeton University Press, vols. I y II.
1950. «Some observations on study design», *Amer. J. Sociol.*, vol. 55, págs. 355-61.

Thistlethwaite, D. L. y Campbell, D. T.
1960. «Regression-discontinuity analysis: An alternative to the ex post facto experiment», *J. Educ. Psychol.*, vol. 51, págs. 309-17.

Thorndike, E. L., McCall, W. A. y Chapman, J. C.
1916. «Ventilation in relation to mental work», *Teach. Coll. Contr. Educ.*, nº 78.

Thorndike, E. L. y Woodworth, R. S.
1901. «The influence of improvement in one mental function upon the efficiency of other functions», *Psychol. Rev.*, vol. 8, págs. 247-61, 384-95, 553-64.

Thorndike, R. L.
1942. «Regression fallacies in the matched groups experiment», *Psychometrika*, vol. 7, págs. 85-102.

Underwood, B. J.
1949. *Experimental psychology*, Nueva York: Appleton-Century-Crofts.
1954. «An analysis of the methodology used to investigate thinking behavior», trabajo leído en la New York University Conference on Human Problem Solving, abril de 1954. (Véase también C. I. Hovland y H. H. Kendler, «The New York University Conference on Human Problem Solving», *Amer. Psychologist*, vol. 10, págs. 64-68.)
1957*a*. «Interference and forgetting», *Psychol. Rev.*, vol. 64, págs. 49-60.
1957*b*. «*Psychological research*, Nueva York: Appleton-Century-Crofts.

Underwood, B. J. y Richardson, J.
1958. «Studies of distributed practice. XVIII. The influence of meaningfulness and intralist similarity of serial nonsense lists», *J. Exp. Psychol.*, vol. 56, págs. 213-19.
1947. U. S. War Department, Information and Education Division, «Opinions about Negro infantry platoons in white companies of seven divisions», en T. M. Newcomb y E. L. Hartley, eds., *Readings in social psychology*, Nueva York: Holt, págs. 542-46. (*Manual de psicología social*, Buenos Aires: Eudeba, 2 vols., 1964.)

Watson, R. I.
1959. *Psychology of the child*, Nueva York: Wiley.

Webb, E. J., Campbell, D. T., Schwartz, R. D. y Sechrest, L.
1966. «Unobtrusive measures: Nonreactive research in the social sciences», Chicago: Rand McNally.

Wilk, M. B. y Kempthorne, O.
1955. «Fixed, mixed, and random models», *J. Amer. Statist. Ass.*, vol. 50, págs. 1144-67.
1956*a*. «Corrigenda», *J. Amer. Statist. Ass.*, vol. 51, pág. 652.
1956*b*. «Some aspects of the analysis of factorial experiments

in a completely randomized desing», *Ann. Math. Statist.*, vol. 27, págs. 950-85.
1957. «Non-additivities in a Latin square design», *J. Amer. Statist. Ass.*, vol. 52, págs. 218-36.

Windle, C.
1954. «Test-retest effect on personality questionnaires», *Educ. Psychol. Measmt.*, vol. 14, págs. 617-33.

Winer, B. J.
1962. *Statistical principles in experimental design*, Nueva York: McGraw-Hill.

Wold, H.
1956. «Causal inference from observational data. A review of ends and means», *J. Royal Statist. Soc.*, sec. A, vol. 119, págs. 28-61.

Wyatt, S., Fraser, J. A. y Stock, F. G. L.
1926. «Fan ventilation in a humid weaving shed», informe nº 37 del Medical Research Council, Industrial Fatigue Research Board, Londres: His Majesty's Stationery Office.

Zeisel, H.
1947. *Say it with figures*, Nueva York: Harper.

Indice onomástico

Indice general

Biblioteca de sociología

Michele Abbate, Libertad y sociedad de masas
Hayward R. Alker, El uso de la matemática en el análisis político
Pierre Ansart, El nacimiento del anarquismo
Pierre Ansart, Las sociologías contemporáneas
David E. Apter, Estudio de la modernización
Peter Bachrach, Crítica de la teoría elitista de la democracia
Brian M. Barry, Los sociólogos, los economistas y la democracia
Reinhard Bendix, Max Weber
Reinhard Bendix, Estado nacional y ciudadanía
Oliver Benson, El laboratorio de ciencia política
Peter L Berger, comp., Marxismo y sociología. Perspectivas desde Europa oriental
Peter L. Berger y *Thomas Luckmann*, La construcción social de la realidad
Norman Birnbaum, La crisis de la sociedad industrial
Hubert M. Blalock, Introducción a la investigación social
Tom Bottomore y Robert Nisbet, comps., Historia del análisis sociológico
Severyn T. Bruyn, La perspectiva humana en sociología
Walter Buckley, La sociología y la teoría moderna de los sistemas
Donald T. Campbell y Julian C. Stanley, Diseños experimentales y cuasi-experimentales en la investigación social
Morris R. Cohen y Ernest Nagel, Introducción a la lógica y al método científico, 2 vols.
Lewis A. Coser, Nuevos aportes a la teoría del conflicto social
Michel Crozier, El fenómeno burocrático, 2 vols.
Michel Crozier, La sociedad bloqueada
David Easton, Esquema para el análisis político
David Easton, comp., Enfoques sobre teoría política
S. N. Eisenstadt, Modernización. Movimientos de protesta y cambio social
Raymond Firth, Elementos de antropología social
Robert W. Friedrichs, Sociología de la sociología
Joseph Gabel, Sociología de la alienación
Anthony Giddens, Las nuevas reglas del método sociológico
Anthony Giddens, La constitución de la sociedad
Erving Goffman, Estigma. La identidad deteriorada
Erving Goffman, Internados. Ensayos sobre la situación social de los enfermos mentales
Erving Goffman, La presentación de la persona en la vida cotidiana
Alvin W. Gouldner, La crisis de la sociología occidental
Daniel Guérin y Ernest Mandel, La concentración económica en Estados Unidos