FIGURA 18.3

Condición B_1 Tratamientos		Condición B_2 Tratamientos	
A_1	A_2	A_1	A_2
M_{A_1}	M_{A_2}	M_{A_1}	M_{A_2}

todas las pruebas estadísticas tienen la misma importancia; las importantes, en efecto, son aquellas directamente relacionadas con los problemas e hipótesis de investigación.

En el presente caso, la hipótesis de interacción [la del inciso 3) anterior] es la más importante, ya que se supone que la discriminación depende del nivel de *habilidad*. Los colegios quizá discriminen a diferentes niveles de habilidad. Como se sugirió antes, las mujeres (A_2) tal vez sean aceptadas más que los hombres (A_1) en el nivel de habilidad más alto (B_1); mientras que quizá sean menos aceptadas en el nivel de habilidad más bajo (B_3). Debería ser evidente que el diseño de investigación no es estático. El tener conocimiento sobre diseño puede ayudar a planear y realizar mejor investigación, y también puede sugerir la comprobación de hipótesis. Y quizá más importante: puede llevar a que uno se dé cuenta de que el diseño de un estudio no es adecuado a las demandas planteadas. ¿Qué significa esta afirmación un tanto peculiar?

Suponga que se formula la hipótesis de interacción como se bosquejó anteriormente, sin saber nada sobre el diseño factorial; en realidad se establece un diseño que consiste de dos experimentos, en uno de los cuales se prueba A_1 contra A_2, bajo la condición B_1. En el segundo experimento se prueba A_1 contra A_2, bajo la condición B_2. El paradigma se vería como el que se muestra en la figura 18.3. (Para simplificar las cosas, únicamente se utilizan dos niveles de B: B_1 y B_2; por lo tanto, el diseño se reduce a uno de 2×2.)

El punto importante a señalar es que no es posible realizar una prueba *adecuada* de la hipótesis con este diseño. A_1 puede probarse contra A_2 bajo las dos condiciones B_1 y B_2 para asegurarse. Pero no es posible saber con claridad y sin ambigüedades, si existe una interacción significativa entre A y B. Aun cuando $M_{A_1} > M_{A_2} \mid B_2$ (M_{A_1} es mayor que M_{A_2}, bajo la condición B_2), como se hipotetizó, el diseño no puede ofrecer una clara posibilidad de confirmación de la interacción hipotetizada, debido a que no se puede obtener información sobre las diferencias entre A_1 y A_2 en los dos niveles de B(B_1 y B_2). Recuerde que una hipótesis de interacción implica, en este caso, que la diferencia entre A_1 y A_2 es distinta en B_1 de lo que es en B_2. En otras palabras, la información tanto de A como de B *juntas en un experimento* es necesaria para probar una hipótesis de interacción. Si los resultados estadísticos de experimentos separados mostraran una diferencia significativa entre A_1 y A_2 en un experimento bajo la condición B_1, y no mostraran diferencias significativas en otro experimento bajo la condición B_2, entonces hay *presunta* evidencia de que la hipótesis de interacción es correcta. Pero no es suficiente contar con presunta evidencia especialmente cuando se sabe que es posible obtener una mejor evidencia.

Suponga que en la figura 18.3, las medias de las casillas fueran, de izquierda a derecha: 30, 30, 40, 30. Tal resultado parecería apoyar la hipótesis de interacción, ya que hay una diferencia significativa entre A_1 y A_2 en el nivel B_2, pero no en el nivel B_1. Pero no puede tenerse la certeza de que esto es así, incluso si la diferencia entre A_1 y A_2 es estadísticamente significativa. La figura 18.4 presenta cómo resultaría esto si se hubiese utilizado un diseño factorial. (Las cifras en las casillas y en los márgenes son medias.) Considerando que los efectos principales, A_1 y A_2; B_1 y B_2, fueran significativos, todavía es posible que la interac-

FIGURA 18.4

	A_1	A_2	
B_1	30	30	30
B_2	40	30	35
	35	30	

ción no sea significativa. A menos que la hipótesis de interacción se pruebe específicamente, la evidencia para determinar la interacción es mera presunción, ya que falta la prueba estadística de la interacción que un diseño factorial proporciona. Debe quedar claro que el conocimiento sobre diseño hubiese mejorado este experimento.

El diseño de investigación como control de la varianza

La principal función técnica del diseño de investigación es *controlar la varianza*. Un diseño de investigación constituye, por así decirlo, un conjunto de instrucciones para que el investigador reúna y analice los datos de cierta forma; por lo tanto, es un mecanismo de control. El principio estadístico que subyace a este mecanismo, como se dijo antes, es: *maximizar la varianza sistemática, controlar la varianza sistemática extraña* y *minimizar la varianza del error*. En otras palabras, se debe *controlar* la varianza.

De acuerdo con este principio, al construir un diseño de investigación eficiente, el investigador intenta: 1) maximizar la varianza de la variable o variables de la hipótesis sustantiva de investigación, 2) controlar la varianza de variables extrañas o "indeseables" que puedan tener un efecto en los resultados experimentales y 3) minimizar la varianza del error o aleatoria, incluyendo los llamados errores de medición. Ahora se verá un ejemplo.

Un ejemplo controversial

La controversia abunda en toda la ciencia y parece ser especialmente rica y variada en las ciencias del comportamiento. Dos controversias han surgido a partir de diferentes teorías del comportamiento y aprendizaje humanos. Los teóricos del reforzamiento han demostrado ampliamente que el reforzamiento positivo puede incrementar el aprendizaje. Sin embargo, como siempre, las cuestiones no son tan simples. El supuesto efecto benéfico de las recompensas externas se ha cuestionado; la investigación ha mostrado que la recompensa extrínseca puede tener una influencia perjudicial en la motivación, interés intrínseco y aprendizaje de los niños. En los años setenta, se publicó una serie de artículos y estudios que mostraban los posibles efectos dañinos del uso de la recompensa. En uno de dichos estudios, Amabile (1979) demostró que la evaluación externa tiene un efecto perjudicial sobre la creatividad artística. Otros estudios incluyen el de Deci (1971) y el de Lepper y Greene (1978). Al mismo tiempo, incluso el principio del reforzamiento en apariencia simple, no es tan simple. Sin embargo, en años recientes han aparecido varios artículos que defienden los efectos positivos de la recompensa (véase Eisenberger y Cameron, 1996; Sharpley, 1988; McCullers, Fabes y Moran, 1987; Bates, 1979).

Existen diversas investigaciones y creencias que indican que los estudiantes universitarios aprenden bien bajo el régimen de lo que se ha llamado *aprendizaje de dominio* (*mastery learning*). De manera sintética, diremos que el "aprendizaje de dominio" consti-

tuye un sistema pedagógico basado en instrucciones personalizadas que requiere que los estudiantes aprendan unidades curriculares hasta alcanzar un criterio de dominio (véase Abbott y Falstrom, 1975; Ross y McBean, 1995; Senemoglu y Fogelman, 1995; Bergin, 1995). Aunque parece existir cierta investigación que apoya la eficacia del aprendizaje de dominio, hay por lo menos un estudio —y es un buen estudio— realizado por Thompson (1980), cuyos resultados indican que los estudiantes a quienes se enseñó con el método de aprendizaje de dominio no fueron mejores que los estudiantes a quienes se enseñó con un enfoque convencional de conferencia, discusión y memorización. Éste es un estudio ejemplar, realizado con controles cuidadosos, durante un largo periodo. El ejemplo que se presenta a continuación estuvo inspirado en el estudio de Thompson. Sin embargo, el diseño y los controles del ejemplo son mucho más simples que los de Thompson. Observe también que Thompson tenía una enorme ventaja: realizó su experimento en un establecimiento militar, lo cual, por supuesto, significa que muchos problemas de control, con frecuencia recalcitrantes en la investigación educativa, se resolvieron fácilmente.

La controversia surge porque los partidarios del aprendizaje de dominio parecen estar fuertemente convencidos de sus virtudes; mientras que los escépticos permanecen incrédulos. ¿Decidirá la investigación el asunto? Es difícil. Pero ahora se verá cómo se podría diseñar un estudio relativamente modesto, capaz de proporcionar por lo menos una respuesta *empírica* parcial.

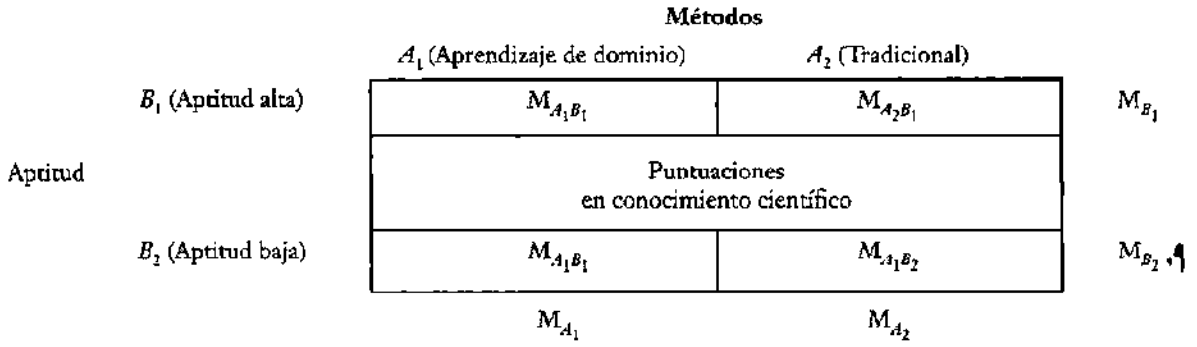
Un investigador educativo decide probar la hipótesis de que el aprovechamiento en ciencia sufre un mayor incremento con un método de aprendizaje de dominio (AD), que con un método tradicional (T). Se ignoran los detalles de los métodos para concentrarse en el diseño de la investigación. Llámese al método de aprendizaje de dominio A_1 y al método tradicional A_2. Los investigadores saben que otras posibles variables independientes ejercen influencia sobre el aprovechamiento: inteligencia, género, antecedentes de clase social, experiencias previas con la ciencia, motivación, etcétera. Existen razones para creer que los dos métodos funcionan de diferente manera con diferentes tipos de estudiantes. Por ejemplo, quizá funcionen de manera diferente con estudiantes con distintos niveles de aptitud escolar. El enfoque tradicional tal vez resulte efectivo con estudiantes con alta aptitud; mientras que el aprendizaje de dominio sea más efectivo con estudiantes con baja aptitud. Llámese B a las aptitudes: aptitud alta es B_1 y aptitud baja es B_2. En este ejemplo la variable aptitud se dicotomizó en los grupos de aptitud alta y baja. Ésta no es la mejor forma de utilizar la variable aptitud; cuando una medida continua se dicotomiza o tricotomiza, se pierde la varianza. En un capítulo posterior se verá que constituye un mejor método respetar el nivel de la medida continua y utilizar una regresión múltiple.

¿Qué tipo de diseño debe establecerse? Para responder es importante etiquetar las variables y saber con claridad cuáles son las preguntas que se formulan. Las variables son:

Variable independiente		Variable dependiente
Métodos	*Aptitud*	*Aprovechamiento en ciencias*
Aprendizaje de dominio, A_1	Aptitud alta, B_1	Puntuaciones de la prueba de ciencia
Tradicional, A_2	Aptitud baja, B_2	

Los investigadores pudieron haber incluido otras variables en el diseño, en especial variables potencialmente influyentes sobre el aprovechamiento: inteligencia general, clase social, género, promedio de preparatoria, por ejemplo. También se podría utilizar la asignación aleatoria para ocuparse de la inteligencia y otras posibles variables independientes de influencia. La medida de la variable dependiente se puede obtener mediante una prueba estandarizada de conocimientos de ciencia.

FIGURA 18.5



Parece que el problema requiere de un diseño factorial. Existen dos razones para esta opción: 1) hay dos variables independientes, 2) es claro que se tiene en mente una hipótesis de interacción, aunque no se haya expresado con tantas palabras. Se cree que los métodos funcionarán de manera diferente con distintos tipos de estudiantes. Se establece la estructura de diseño que se representa en la figura 18.5.

Observe que todas las medias marginales y de casilla han sido etiquetadas de forma apropiada. Note también que hay una *variable activa*, métodos; y una *variable atributo*, aptitudes. Quizá recuerde del capítulo 3 que una *variable activa* es una variable experimental o manipulada; una *variable atributo* es una variable medida o una variable que es una característica de personas o grupos; por ejemplo, inteligencia, clase social y ocupación (gente); así como cohesión, productividad y atmósfera restrictiva-permisiva (organizaciones, grupos, etcétera). Todo lo que puede hacerse es categorizar a los participantes como con aptitud alta y aptitud baja, y asignarlos de acuerdo con ello a B_1 y B_2. Sin embargo, es posible asignar a los estudiantes aleatoriamente a A_1 y A_2, los grupos de los métodos. Esto se realiza en dos etapas: 1) los estudiantes de B_1 (aptitud alta) se asignan aleatoriamente a A_1 y A_2, y 2) los estudiantes de B_2 (aptitud baja) se asignan aleatoriamente a A_1 y A_2. Al aleatorizar así a los participantes se puede suponer que antes de que empiece el experimento, los estudiantes en A_1 son aproximadamente iguales a los estudiantes en A_2, en todas las características posibles.

El interés aquí radica en los diferentes papeles de la varianza en el diseño de investigación y en el principio de la varianza. Antes de continuar, al principio de la varianza se le llamará "maxmincon" para su fácil referencia. El origen del nombre es evidente: maximizar la varianza sistemática en estudio; controlar la varianza sistemática extraña y minimizar la varianza del error, con dos sílabas invertidas por eufonía.

Antes de ilustrar la aplicación del principio maxmincon en el presente ejemplo, debe discutirse un punto importante. Siempre que se hable de varianza hay que estar seguro de saber de qué tipo de varianza se habla. Se habla de la varianza de los métodos, de inteligencia, de género, de tipo de hogar, etcétera; parecería que se refiere a la varianza de la variable independiente, lo cual es verdad y, a la vez, no. Siempre se refiere a la *varianza de la variable dependiente, y a la varianza de las medidas de la variable dependiente*, después de que se realizó el experimento. Esto no es cierto en los llamados estudios correlacionales, donde al decir "la varianza de la variable independiente" significa justamente eso. Al correlacionar dos variables, se estudian "directamente" las varianzas de las variables dependiente e independiente. La referencia "varianza de la variable independiente" surge del hecho de que,

mediante la manipulación y el control de las variables independientes, presumiblemente se ejerce influencia sobre la varianza de la variable dependiente. Dicho de manera algo imprecisa, se "hace" que las medidas de la variable dependiente se comporten o varíen como un supuesto resultado de la manipulación y el control de las variables independientes. En un experimento se analizan las medidas de la variable dependiente y, a partir del análisis, se infiere que las varianzas presentes en la varianza total de las medidas de la variable dependiente se deben a la manipulación y control de las variables independientes y no al error. Ahora regresemos al principio en cuestión.

Maximización de la varianza experimental

La preocupación más obvia del investigador, aunque no necesariamente la más importante, consiste en maximizar la llamada *varianza experimental.* Dicho término se introduce para facilitar discusiones subsecuentes y, en general, tan sólo se refiere a la varianza de la variable dependiente, debida a la influencia ejercida por la variable independiente o variables de la hipótesis sustantiva. En este caso en particular, la varianza experimental es la varianza en la variable dependiente, presumiblemente debida a los métodos A_1 y A_2, y a los niveles de aptitud B_1 y B_2. Aunque la varianza experimental puede tomarse para hacer referencia únicamente a la varianza debida a la variable manipulada o *activa*, como los métodos, también se pueden considerar las variables *atributo* como inteligencia, género y, en este caso, aptitud, como variables experimentales. Una de las principales tareas de un experimentador consiste en maximizar esta varianza. Los métodos deben "separarse" lo más posible para hacer a A_1 y A_2 (y A_3, A_4, etcétera, si están en el diseño) tan diferentes como sea posible.

Si la variable independiente no varía de manera sustancial, hay poca posibilidad de separar su efecto de la varianza total de la variable dependiente. Es necesario dar a la varianza de una relación la oportunidad de mostrarse, de separarse, por así decirlo, de la varianza total, la cual es un compuesto de varianzas debidas a numerosas fuentes y al azar. Teniendo presente este subprincipio del principio maxmincon, se puede declarar un precepto de investigación: *diseñar, planear y conducir la investigación de tal forma que las condiciones experimentales sean tan diferentes como sea posible.* Existen, por supuesto, excepciones a este subprincipio, pero probablemente sean poco comunes. Puede ser que un investigador desee estudiar los efectos de pequeñas gradaciones de, digamos, incentivos motivacionales sobre el aprendizaje de algún tema. Aquí no se buscaría que las condiciones experimentales fueran lo más diferentes posibles; aun así, debería asegurarse de que varían un poco o no habría una varianza resultante discernible en la variable dependiente.

En el presente ejemplo de investigación, este subprincipio se refiere a que el investigador debe realizar un esfuerzo para hacer a A_1 y A_2, los métodos de aprendizaje de dominio y el tradicional, tan diferentes como sea posible. Después, B_1 y B_2 también deben ser tan diferentes como sea posible, en la dimensión de aptitud. Este último problema en esencia es uno de medición, como se verá en un capítulo posterior. En un experimento, el investigador es como un titiritero que hace que los títeres de la variable independiente hagan lo que él quiere. Sostiene los hilos de los títeres A_1 y A_2 con la mano derecha; y los hilos de los títeres B_1 y B_2, con la mano izquierda. (Se supone que una mano no ejerce influencia sobre la otra, es decir, las manos deben ser independientes.) Los títeres A_1 y A_2 se ponen a bailar por separado, al igual que los títeres B_1 y B_2; entonces, el investigador presta atención a la audiencia (la variable dependiente) para observar y medir el efecto de las manipulaciones. Si tiene éxito al hacer bailar a A_1 y A_2 por separado, y si existe una relación entre A y la variable dependiente, y, si por ejemplo, separar A_1 de A_2 es gracioso,

la reacción de la audiencia debería ser una carcajada. El investigador incluso puede notar que sólo consigue risas cuando A_1 y A_2 bailan por separado y, al mismo tiempo, B_1 y B_2 bailan por separado (interacción de nuevo).

Control de variables extrañas

El control de variables extrañas se refiere a minimizar, anular o aislar las influencias de aquellas variables independientes extrañas a los propósitos del estudio. Hay tres formas de controlar las variables extrañas. El primero es el más sencillo, si es posible realizarlo: eliminar la variable como tal. Si existe preocupación sobre la inteligencia como un posible factor contribuyente en estudios de aprovechamiento, su efecto sobre la variable dependiente virtualmente puede ser eliminado utilizando participantes con un solo nivel de inteligencia, digamos puntuaciones de inteligencia dentro del rango de 90 a 110. Si se estudia el aprovechamiento, y el origen racial es un posible factor contribuyente a la varianza del aprovechamiento, se elimina utilizando únicamente miembros de una raza. El principio es: *eliminar el efecto de una variable independiente que posiblemente influya sobre la variable dependiente, es decir, elegir a los participantes de manera que sean lo más homogéneos posible en esa variable independiente.*

Este método para controlar la varianza indeseable o extraña es muy efectivo. Si se selecciona solamente un género para un experimento, entonces se tiene la seguridad de que el género no sea una variable independiente contribuyente. Pero entonces se pierde el poder de generalización; por ejemplo, no es factible hablar sobre la relación estudiada respecto a las niñas si únicamente se utilizan niños en el experimento. Si se restringe el rango de inteligencia, entonces solamente se analiza dicho rango restringido. ¿Es posible que la relación, si se descubre una, sea inexistente o muy distinta con niños de alta inteligencia o con niños de baja inteligencia? Simplemente no se sabe; tan sólo se puede conjeturar o suponer.

La segunda forma para controlar la varianza extraña es a través de la aleatorización. Ésta es la mejor manera, en el sentido de que es posible tener el pastel y también comer un poco de él. En teoría, la aleatorización es el único método para controlar todas las variables extrañas posibles. Otra forma de expresarlo es: si se logra una aleatorización adecuada, entonces los grupos experimentales pueden ser considerados estadísticamente iguales en todas las formas posibles. Por supuesto que esto no quiere decir que los grupos sean iguales en todas las variables posibles. Ya se sabe que, por azar, los grupos pueden ser desiguales; pero con la aleatorización adecuada, la probabilidad de que sean iguales es mayor que la probabilidad de que no lo sean. Por tal razón, el control de la varianza extraña por medio de la aleatorización es un poderoso método de control. Todos los otros métodos dejan abiertas muchas posibilidades de desigualdad. Si se aparean los grupos respecto a la inteligencia, quizá se logre exitosamente la igualdad estadística en inteligencia (al menos en aquellos aspectos de inteligencia que se miden), pero se puede sufrir de desigualdad en otras variables independientes significativamente influyentes como aptitud, motivación y clase social. Un precepto que surge a partir de este poder igualador de la aleatorización es: *siempre que sea posible hacerlo, asigne a los sujetos a los grupos y a las condiciones experimentales de manera aleatoria, y asigne las condiciones y otros factores a los grupos experimentales de manera aleatoria.*

El tercer método para controlar una variable extraña es incluirla en el diseño como una variable independiente. Por ejemplo, suponga que en el experimento discutido anteriormente se fuera a controlar el género y que se considerara inoportuno o imprudente eliminarlo. Se podría añadir una tercera variable al diseño: el género. A menos que se estuviera interesado en la diferencia real entre los géneros respecto a la variable depen-

diente, o se deseara estudiar la interacción entre una o dos de las otras variables y el género, es poco probable que se utilice esta forma de control. Se puede desear información del tipo antes mencionado y también desear controlar el género. En tal caso sería deseable añadirlo al diseño como una variable. El punto es que incorporar una variable a un diseño experimental "controla" la variable, ya que, entonces, resulta posible extraer de la varianza total de la variable dependiente, la varianza debida a la variable. (En el caso anterior se trataría de la varianza "entre géneros".)

Tales consideraciones llevan a otro principio: *una variable extraña puede ser controlada al incorporarla al diseño de investigación como una variable atributo, logrando así control y proporcionando información adicional de investigación sobre el efecto de la variable sobre la variable dependiente y sobre su posible interacción con otras variables independientes.*

La cuarta forma para controlar la varianza extraña consiste en aparear a los participantes. El principio de control detrás del apareamiento es el mismo que aquel para cualquier otra forma de control: el control de la varianza. El apareamiento es similar —de hecho podría llamarse un corolario— al principio del control de la varianza de una variable extraña mediante su incorporación al diseño. El principio básico consiste en dividir una variable en dos o más partes en un diseño factorial, digamos en inteligencia alta y baja, y después aleatorizarla dentro de cada nivel, como se describió antes. El apareamiento es un caso especial de este principio. Sin embargo, en lugar de dividir a los participantes en dos, tres o cuatro partes, se dividen en $N/2$ partes; donde N es el número de participantes involucrados; de esta manera se incorpora al diseño el control de la varianza.

Con el uso del método de apareamiento pueden surgir varios problemas. En primer lugar, la variable respecto a la cual se aparean los participantes debe estar sustancialmente relacionada con la variable dependiente o el apareamiento es una pérdida de tiempo; aun peor, puede causar confusión. Además, el apareamiento tiene limitaciones severas. Si se intenta aparear, digamos, más de dos variables, o incluso más de una, se pierden participantes. Es difícil encontrar participantes apareados en más de dos variables. Por ejemplo, si se decide aparear a los sujetos en cuanto a inteligencia, género y clase social, se puede tener éxito al aparear las dos primeras variables, pero se puede fracasar al intentar encontrar pares que sean bastante iguales en las tres variables. Añádase una cuarta variable y el problema se torna difícil, incluso con frecuencia imposible de resolver.

Sin embargo, no se tire al bebé con el agua del baño. Cuando existe una correlación sustancial entre la variable o variables apareadas y la variable dependiente (mayor que .50 o .60), entonces el apareamiento reduce el término del error y aumenta la precisión de un experimento, un resultado deseable. Si se utilizan los mismos participantes con diferentes tratamientos experimentales —llamados medidas repetidas o diseño de bloque aleatorizado—, entonces se tienen poderosos controles de la varianza. ¿Cómo se puede realizar un mejor apareamiento en todas las variables posibles que apareando un sujeto consigo mismo? Por desgracia, otras consideraciones negativas generalmente impiden dicha posibilidad. Debe enfatizarse con vigor que el apareamiento de cualquier tipo no sustituye la aleatorización. Si se aparea a los participantes, *entonces ellos deben ser asignados aleatoriamente a los grupos experimentales*. A través de un procedimiento aleatorio, como lanzar una moneda o utilizar números pares o nones aleatorios, los miembros de los pares apareados son asignados a los grupos experimental y control. Si los mismos participantes son sometidos a todos los tratamientos, entonces el orden de los tratamientos debe asignarse aleatoriamente. Esto añade control de aleatorización al apareamiento, o control de medidas repetidas.

Un principio sugerido por el presente análisis es: *cuando una variable apareada se correlaciona sustancialmente con la variable dependiente, el apareamiento, como forma de control de la varianza, resulta útil y deseable*. Sin embargo, antes de realizar un apareamiento se

deben sopesar cuidadosamente sus ventajas y desventajas en una situación de investigación particular. La aleatorización completa o el análisis de covarianza pueden ser mejores métodos de control de la varianza.

Otra forma de control, el control estadístico, se analizó en capítulos previos; pero uno o dos comentarios son pertinentes en este momento. Los métodos estadísticos constituyen, por así decirlo, formas de control en el sentido de que aíslan y cuantifican las varianzas. Pero el control estadístico es inseparable de otras formas de control de diseño. Por ejemplo, si se utiliza el apareamiento, debe usarse una prueba estadística apropiada, de otro modo se perderá el efecto del apareamiento y, por lo tanto, el control.

Minimización de la varianza del error

La *varianza del error* es la variabilidad de las medidas debidas a fluctuaciones aleatorias, cuya característica básica es que son autocompensatorias, es decir, en un momento varían de una forma, luego de otra, a veces positiva, a veces negativa, a veces hacia arriba, a veces hacia abajo. Los errores aleatorios tienden a equilibrarse entre sí, de tal manera que su media es cero.

Existen varios determinantes de la varianza del error, por ejemplo, factores asociados con las diferencias individuales entre los participantes. Por lo común, a esta varianza debida a diferencias individuales se le llama "varianza sistemática". Pero cuando dicha varianza no puede identificarse ni controlarse, debe agruparse con la varianza del error. Puesto que muchos determinantes interactúan y tienden a cancelarse entre sí (o al menos se supone que lo hacen), la varianza del error tiene estas características aleatorias.

Otra fuente de la varianza del error es aquella asociada con los llamados errores de medición: variación de las respuestas de un ensayo a otro, adivinación, inatención momentánea, fatiga ligera temporal, fallas de la memoria, estados emocionales transitorios de los participantes, etcétera.

Minimizar la varianza del error tiene dos aspectos principales: 1) la reducción de los errores de medición a través de condiciones controladas y 2) un aumento en la confiabilidad de las medidas. Mientras menor sea el control de las condiciones de un experimento, mayor será la influencia de los muchos determinantes de la varianza del error. Ésta es una de las razones para establecer con cuidado condiciones experimentales controladas. En estudios bajo condiciones de campo, por supuesto, dicho control se vuelve difícil; aun así, deben realizarse esfuerzos constantes para disminuir los efectos de los múltiples determinantes de la varianza del error. Esto puede hacerse, en parte, dando instrucciones claras y específicas a los participantes, y excluyendo de la situación experimental los factores que sean extraños al propósito de la investigación.

Incrementar la confiabilidad de las medidas implica reducir la varianza del error. Aunque en posteriores capítulos se incluyen análisis más completos sobre el tema, de momento diremos que la confiabilidad se considera como la precisión de un conjunto de puntuaciones. Hasta el punto en que las puntuaciones no fluctúen aleatoriamente, son confiables. Imagine un instrumento de medición no confiable por completo; dicho instrumento no permite predecir el desempeño futuro de los individuos; en un momento brinda un ordenamiento de los rangos para una muestra de participantes y un ordenamiento diferente de los rangos en otro momento. Con un instrumento de este tipo no sería posible identificar ni extraer las varianzas sistemáticas, debido a que las puntuaciones generadas por el instrumento serían como números en una tabla de números aleatorios, que es el caso extremo. Ahora imaginen cantidades diferentes de confiabilidad y de no confiabilidad en las medidas de la variable dependiente. Cuanto más confiables sean las medidas, mejor se

podrán identificar y extraer las varianzas sistemáticas y menor será la varianza del error en relación con la varianza total.

Otra razón para reducir la varianza del error tanto como sea posible, es darle la oportunidad a la varianza sistemática para que se muestre a sí misma, lo cual no puede hacerse si la varianza del error y, por lo tanto, el término del error, son demasiado grandes. Si existe una relación, se busca descubrirla. Una forma de descubrir la relación consiste en encontrar diferencias significativas entre las medias. Pero si la varianza del error es relativamente grande debido a errores de medición no controlados, la varianza sistemática —llamada antes varianza "entre"— no tendrá la oportunidad de aparecer. Por lo tanto, aunque existente, la relación probablemente no será detectada.

El problema de la varianza del error puede expresarse con claridad en forma matemática: recuerde la siguiente ecuación:

$$V_t = V_e + V_r$$

donde V_t es la varianza total en un conjunto de medidas; V_e es la varianza entre grupos, la varianza presumiblemente debida a la influencia de las variables experimentales; y V_r es la varianza del error (en el análisis de varianza, la varianza dentro de grupos y la varianza residual). En efecto, a mayor valor de V_t, menor deberá ser V_r, con una cantidad dada de V_t.

Considere la siguiente ecuación: $F = V_e/V_r$. Para que el numerador de la fracción a la derecha sea evaluado con precisión respecto a una desviación significativa de las expectativas por el azar, el denominador debe ser una medida exacta del error aleatorio.

Un ejemplo familiar sirve para aclarar esto. Recuerde que en la discusión sobre el análisis de varianza factorial y sobre el análisis de varianza de grupos correlacionados, se habló sobre la varianza debida a las diferencias individuales presentes en las medidas experimentales. Se afirmó que, mientras que la aleatorización adecuada puede igualar efectivamente a los grupos experimentales, habrá varianza en las puntuaciones debida a las diferencias individuales; por ejemplo, diferencias debidas a la inteligencia, aptitud, etcétera. Ahora, en algunas situaciones, estas diferencias individuales pueden ser bastante grandes. Si lo son, entonces la varianza del error y, en consecuencia, el denominador de la ecuación F anterior, serán "demasiado grandes" en relación con el numerador; es decir, las diferencias individuales habrán sido aleatoriamente dispersadas entre, digamos, dos, tres o cuatro grupos experimentales. Aun así, son fuentes de varianza y, como tales, inflarán la varianza dentro de los grupos o la residual, es decir, el denominador de la ecuación anterior.

RESUMEN DEL CAPÍTULO

1. Los diseños de investigación son planes y estructuras utilizados para responder preguntas de investigación.
2. Los diseños de investigación tienen dos propósitos básicos: (i) proporcionar respuestas a preguntas de investigación, y (ii) controlar la varianza.
3. Los diseños de investigación funcionan en conjunto con las hipótesis de investigación para generar una respuesta confiable y válida.
4. Los diseños de investigación también pueden indicar qué prueba estadística emplear para analizar los datos recolectados a partir de ese diseño.
5. Hablar de controlar la varianza, puede referirse a una o más de tres cuestiones:
   - maximizar la varianza sistemática
   - controlar la varianza extraña
   - minimizar varianza del error

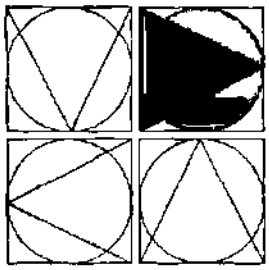
6. Para maximizar la varianza sistemática debe tenerse una variable independiente, cuyos niveles sean muy distintos entre sí.
7. Para controlar la varianza extraña, el investigador necesita eliminar los efectos de una variable independiente potencial sobre la variable dependiente, lo cual puede hacerse por medio de:

   - mantener constante la variable independiente; por ejemplo, si se sabe que el género tiene un efecto potencial, entonces puede mantenerse constante al realizar el estudio incluyendo sólo un género (por ejemplo mujeres).
   - la aleatorización, que se refiere a elegir participantes de manera aleatoria y después asignar aleatoriamente a cada grupo de participantes a las condiciones de tratamiento (niveles de la variable independiente).
   - incorporar la variable extraña al diseño, convirtiéndola en una variable independiente.
   - apareando a los participantes; este método de control puede resultar difícil en ciertas situaciones; un investigador nunca podrá estar muy seguro de que se realizó un apareamiento exitoso en todas las variables importantes.

8. Minimizar la varianza del error incluye la medición de la variable dependiente. Al reducir el error de medición se reduce la varianza del error. El incremento en la confiabilidad de la medición también conlleva una reducción de la varianza del error.

SUGERENCIAS DE ESTUDIO

1. Hemos notado que el diseño de investigación tiene el propósito de obtener respuestas a las preguntas de investigación y controlar la varianza. Explique en detalle qué significa esta afirmación. ¿Cómo controla la varianza un diseño de investigación? ¿Por qué un diseño factorial debería controlar más varianza que un diseño de un factor? ¿Cómo es que un diseño que utiliza participantes apareados o medidas repetidas de los mismos participantes controla la varianza? ¿Cuál es la relación entre las preguntas de investigación, las hipótesis de investigación y un diseño de investigación? Invente un problema de investigación para ilustrar sus respuestas a estas preguntas (o utilice un ejemplo del texto).
2. Sir Ronald Fisher (1951), el inventor del análisis de varianza, en uno de sus libros dijo que debe aclararse que la hipótesis nula nunca se confirma o establece; pero que es posible refutarla en el curso de la experimentación. Se puede decir que todo experimento existe para dar a los hechos la oportunidad de refutar la hipótesis nula. Ya sea que usted esté de acuerdo o no con la afirmación de Fisher, ¿qué piensa que quiso decir con ello? Para estructurar su respuesta, recuerde el principio maxmincon y las pruebas F y t.



CAPÍTULO 19

DISEÑOS INADECUADOS Y CRITERIOS PARA EL DISEÑO

- ENFOQUES EXPERIMENTAL Y NO EXPERIMENTAL
- SIMBOLOGÍA Y DEFINICIONES
- DISEÑOS DEFECTUOSOS
  - **Medición, historia, maduración**
  - **El efecto de regresión**
- CRITERIOS DEL DISEÑO DE INVESTIGACIÓN
  - **¿Responder preguntas de investigación?**
  - **Control de variables independientes extrañas**
  - **Posibilidad de generalización**
  - **Validez interna y externa**

Todas las creaciones de las disciplinas de los seres humanos tienen forma. La arquitectura, la poesía, la música, la pintura, las matemáticas, la investigación científica, todas tienen forma. La gente pone gran énfasis en el contenido de sus creaciones, frecuentemente sin darse cuenta de que sin una estructura fuerte, no importa cuán rico y significativo sea el contenido, las creaciones pueden resultar débiles y estériles.

Lo mismo sucede con la investigación científica. El científico requiere de una forma viable y flexible con la cual expresar las metas científicas. Sin contenido —sin una buena teoría, buenas hipótesis, buenos problemas— el diseño de investigación está vacío. Pero sin forma, sin una estructura concebida y creada adecuadamente para el propósito de la investigación, pueden lograrse pocas cosas valiosas. De hecho, no es exagerado afirmar que muchos de los fracasos en la investigación del comportamiento han sido fallas en las formas de disciplina e imaginación.

El enfoque principal de este capítulo son los diseños de investigación inadecuados. Tales diseños han sido tan comunes que requieren analizarse. Más importante aún, es tener presente que el estudiante debe ser capaz de reconocerlos y de entender por qué son inadecuados. Este enfoque negativo tiene una virtud: el estudio de las deficiencias obliga a preguntarse por qué algo es deficiente, lo que a su vez centra la atención en los criterios

6. Para maximizar la varianza sistemática debe tenerse una variable independiente, cuyos niveles sean muy distintos entre sí.
7. Para controlar la varianza extraña, el investigador necesita eliminar los efectos de una variable independiente potencial sobre la variable dependiente, lo cual puede hacerse por medio de:

   - mantener constante la variable independiente; por ejemplo, si se sabe que el género tiene un efecto potencial, entonces puede mantenerse constante al realizar el estudio incluyendo sólo un género (por ejemplo mujeres).
   - la aleatorización, que se refiere a elegir participantes de manera aleatoria y después asignar aleatoriamente a cada grupo de participantes a las condiciones de tratamiento (niveles de la variable independiente).
   - incorporar la variable extraña al diseño, convirtiéndola en una variable independiente.
   - apareando a los participantes; este método de control puede resultar difícil en ciertas situaciones; un investigador nunca podrá estar muy seguro de que se realizó un apareamiento exitoso en todas las variables importantes.

8. Minimizar la varianza del error incluye la medición de la variable dependiente. Al reducir el error de medición se reduce la varianza del error. El incremento en la confiabilidad de la medición también conlleva una reducción de la varianza del error.

SUGERENCIAS DE ESTUDIO

1. Hemos notado que el diseño de investigación tiene el propósito de obtener respuestas a las preguntas de investigación y controlar la varianza. Explique en detalle qué significa esta afirmación. ¿Cómo controla la varianza un diseño de investigación? ¿Por qué un diseño factorial debería controlar más varianza que un diseño de un factor? ¿Cómo es que un diseño que utiliza participantes apareados o medidas repetidas de los mismos participantes controla la varianza? ¿Cuál es la relación entre las preguntas de investigación, las hipótesis de investigación y un diseño de investigación? Invente un problema de investigación para ilustrar sus respuestas a estas preguntas (o utilice un ejemplo del texto).
2. Sir Ronald Fisher (1951), el inventor del análisis de varianza, en uno de sus libros dijo que debe aclararse que la hipótesis nula nunca se confirma o establece; pero que es posible refutarla en el curso de la experimentación. Se puede decir que todo experimento existe para dar a los hechos la oportunidad de refutar la hipótesis nula. Ya sea que usted esté de acuerdo o no con la afirmación de Fisher, ¿qué piensa que quiso decir con ello? Para estructurar su respuesta, recuerde el principio maximincon y las pruebas F y t.

utilizados para juzgar tanto las adecuaciones como las inadecuaciones. Así, el estudio de diseños inadecuados conduce al estudio de los criterios del diseño de investigación. También se aprovecha la ocasión para describir el sistema simbológico a utilizar, así como para identificar una distinción importante entre investigación experimental y no experimental.

Enfoques experimental y no experimental

La discusión sobre el diseño se inicia por medio de una distinción importante: aquella entre los enfoques experimental y no experimental de la investigación. De hecho, tal distinción es tan importante que un capítulo separado (capítulo 23) se dedicará a dicho tema. Un *experimento* es una investigación científica donde un investigador manipula y controla una o más variables independientes y observa la(s) variable(s) dependiente(s) para determinar si hay variación concomitante a la manipulación de las variables independientes. Un *diseño experimental*, entonces, es aquel en el que el investigador *manipula* por lo menos una variable independiente. En un capítulo anterior se analizó brevemente el estudio clásico de Hurlock (1925), quien manipuló incentivos para producir diferentes cantidades de retención. En el estudio de Walster, Cleary y Clifford (1970) (capítulo 18), se manipularon género, raza y niveles de habilidad para estudiar sus efectos en la aceptación universitaria: las solicitudes enviadas a las universidades difirieron en las descripciones de los solicitantes como hombre-mujer; blanco-negro; y niveles altos, medios y bajos de habilidad.

En la investigación no experimental no es posible manipular las variables o asignar aleatoriamente a los participantes o tratamientos debido a que la naturaleza de las variables es tal que imposibilita su manipulación. Los participantes llegan al investigador con sus características distintivas intactas, por así decirlo. Vienen con su "ya presente" sexo, inteligencia, nivel ocupacional, creatividad o aptitud. Wilson (1996) utilizó un diseño no experimental para estudiar la legibilidad, contenido étnico y sensibilidad cultural del material educativo sobre pacientes, utilizado por los enfermeros del departamento local de salud y centros comunitarios de salud. En dicho caso el material ya existía; no hubo asignación o selección aleatorias. Edmondson (1996) también usó un diseño no experimental para comparar el número de errores de medicación cometidos por enfermeros, médicos y boticarios en ocho unidades hospitalarias de dos hospitales urbanos de enseñanza. Edmondson no eligió de manera aleatoria estas unidades u hospitales, ni a los profesionales médicos. De la misma forma, en muchas áreas de investigación, por desgracia no es posible realizar asignaciones aleatorias, como se verá más adelante. Aunque la investigación experimental y la no experimental difieren en estos aspectos cruciales, comparten características estructurales y de diseño que se indicarán en éste y subsecuentes capítulos. Además, su propósito básico es el mismo: estudiar relaciones entre fenómenos. Su lógica científica también es la misma: obtener evidencia empírica para realizar proposiciones condicionales de la forma si p, entonces q. En algunos campos de las ciencias sociales y del comportamiento, las estructuras no experimentales son inevitables. Keith (1988) afirma que muchos estudios conducidos por psicólogos escolares son de naturaleza no experimental. Los investigadores de psicología escolar, así como muchos en psicología educativa deben trabajar dentro de una estructura práctica. Muchas veces, escuelas, salones de clase e incluso estudiantes son dados al investigador "como son". Stone-Romero, Weaver y Glenar (1995) sintetizaron casi 20 años de artículos del *Journal of Applied Psychology*, respecto al uso de diseños de investigación experimentales y no experimentales.

El ideal de la ciencia es el experimento controlado. Excepto, quizás, en investigación taxonómica —aquella que tiene el propósito de descubrir, clasificar y medir fenómenos naturales y los factores que subyacen a dichos fenómenos— donde el modelo de ciencia

deseado es el experimento controlado. Puede ser difícil para muchos estudiantes aceptar esta afirmación más bien categórica, puesto que su lógica aún no es aparente. Anteriormente se indicó que la meta principal de la ciencia era descubrir relaciones entre fenómenos; entonces, ¿por qué dar prioridad al experimento controlado? ¿No existen otros métodos para descubrir relaciones? Sí, por supuesto que existen. Sin embargo, la principal razón para la preeminencia del experimento controlado es que los investigadores pueden tener más confianza en que las relaciones que ellos estudian son las relaciones que creen que son. La razón no es difícil de ver: ellos estudian las relaciones bajo las condiciones más cuidadosamente controladas de indagación que se conocen. Así, la virtud única y abrumadoramente importante del estudio experimental es el control. En un estudio experimental perfectamente controlado, el investigador puede confiar en que la manipulación de la variable independiente es lo que afectó a la variable dependiente, y nada más. En resumen, un estudio experimental perfectamente conducido es más confiable que un estudio no experimental perfectamente conducido. La razón de ello debe volverse más obvia conforme se avance en el estudio del diseño de investigación.

Simbología y definiciones

Antes de discutir los diseños inadecuados, resulta necesario explicar la simbología utilizada en estos capítulos. X se utiliza para definir una variable (o variables) independiente que es *experimentalmente manipulada*. X_1, X_2, X_3, etcétera, representan las variables independientes 1, 2, 3, etcétera, aunque por lo común se utiliza la X sola, aun cuando pueda significar más de una variable independiente. (También se utiliza X_1, X_2, ..., para representar las particiones de una variable independiente; pero la diferencia siempre se hará clara.) El símbolo (X) indica que la variable independiente no está *manipulada* —no está bajo el control directo del investigador, sino que es *medida o imaginada*—. La variable dependiente es Y: Y_a es la variable independiente *antes* de la manipulación de X, y Y_d es la variable dependiente *después* de la manipulación de X. Con $\sim X$ se toma prestado el signo de negación de la teoría de conjuntos: $\sim X$ ("no-X") para indicar que la variable experimental (la variable independiente X) *no* está manipulada. [Nota: (X) es una variable no manipulable y $\sim X$ es una variable manipulable que no está manipulada.] El símbolo (A) se utilizará para la asignación aleatoria de los participantes a los grupos experimentales y para la asignación aleatoria de los tratamientos experimentales a los grupos experimentales.

La explicación dada respecto a $\sim X$ no es muy precisa puesto que en algunos casos $\sim X$ puede representar un aspecto diferente del tratamiento X, más que la simple ausencia de tratamiento. En el lenguaje científico usado antes, el grupo experimental era el grupo al que se le daba el llamado tratamiento experimental, X; mientras que el grupo control no lo recibía, $\sim X$. Para los propósitos del texto, sin embargo, $\sim X$ será suficiente, especialmente si se entiende el significado generalizado de *control* explicado antes. Entonces, un *grupo experimental* es un grupo de participantes que reciben algún aspecto o tratamiento de X. En la comprobación de la hipótesis de frustración-agresión, el grupo experimental es aquel a cuyos participantes se les induce frustración sistemáticamente. En contraste, el grupo control es aquel al que "no" se le da tratamiento.

En la investigación multivariada moderna es necesario expandir estos conceptos. En esencia no cambian, sino que se expanden. Como se ha visto, es muy posible tener más de un grupo experimental. No sólo son posibles diferentes grados de manipulación de la variable independiente, sino que con frecuencia son deseables, e incluso imperativos. Además, es posible incluir más de un grupo control, afirmación que de entrada parece absurda. ¿Cómo es posible tener diferentes grados de "no" tratamiento experimental? Esto

sucede porque el concepto de *control* se generaliza. Cuando hay más de dos grupos, y cuando cualquier par de éstos es tratado de manera diferente, uno o más grupos sirven como "controles" para los otros. Recuerde que el control se refiere siempre al control de varianza. Con dos o más grupos tratados de forma diferente, la varianza es generada por la manipulación experimental. Así, el concepto tradicional de X y $\sim X$ (tratamiento y no tratamiento) se generaliza a X_1, X_2, X_3,..., X_k, formas o grados diferentes de tratamiento.

Si X está entre paréntesis (X), significa que el investigador "se imagina" la manipulación de X, o supone que X ocurrió y que se trata de la X de la hipótesis. También puede significar que X está siendo medida y no manipulada. En realidad aquí se está señalando lo mismo de diferente forma; el contexto del análisis debería dejar en claro la distinción. Suponga que un sociólogo estudia la delincuencia y la hipótesis de frustración-agresión. El investigador observa la delincuencia, Y, e imagina que los participantes delincuentes sufrieron frustración en sus primeros años, o (X). Todos los diseños no experimentales tendrán (X); entonces, (X) por lo común representa una variable independiente *que no está bajo el control experimental del investigador.*

Un punto más: en general cada diseño en este capítulo tendrá una forma *a* y una *b*. La forma *a* será la forma experimental, o aquella en la cual se manipula X. La forma *b* será la forma no experimental, en la cual X no está bajo el control del investigador, o (X). Obviamente también es posible ($\sim X$).

Diseños defectuosos

Existen cuatro (o más) diseños de investigación inadecuados que con frecuencia se han usado —y aún se utilizan ocasionalmente— en la investigación del comportamiento. Los defectos de los diseños conducen a un control pobre de las variables independientes. A continuación se enumera cada uno de estos diseños, se le da un nombre, se esquematiza su estructura y después se analiza.

Diseño 19.1: De un grupo		
(*a*) X	*Y*	(Experimental)
(*b*) (X)	*Y*	(No experimental)

El diseño 19.1 ha sido llamado "estudio de caso de un disparo", un nombre pertinente asignado por Campbell y Stanley (1963). La forma (*a*) del diseño es experimental y la forma (*b*) es no experimental. Un ejemplo de investigación con un diseño de la forma (*a*): el cuerpo docente de una facultad instituye un nuevo currículum y busca evaluar sus efectos. Después de un año, se mide Y, el aprovechamiento de los estudiantes. Se concluye, digamos, que el aprovechamiento se ha incrementado con el nuevo programa. Con un diseño de este tipo, la conclusión es débil. El diseño 19.1(*b*) es la forma no experimental del diseño de un grupo. Se estudia Y, el resultado; y X se supone o imagina. Un ejemplo sería el estudio de la delincuencia al analizar el pasado de un grupo de delincuentes juveniles para identificar los factores que probablemente los hayan conducido a su comportamiento antisocial. El método es problemático debido a que pueden confundirse los factores (variables). Cuando el efecto de dos o más factores (variables) no puede separarse, los resultados se vuelven difíciles de interpretar; cualquier número de posibles explicaciones serían plausibles.

Desde el punto de vista científico, el diseño 19.1 carece de valor. Virtualmente no hay control de otras posibles influencias sobre el resultado. Como Campbell señaló (1957), el

mínimo de información científica útil requiere de por lo menos una comparación formal. El ejemplo del currículum requiere, *por lo menos*, una comparación entre el grupo que experimentó el nuevo currículum con otro que no lo haya experimentado. El supuesto efecto del nuevo currículum, digamos tal y cual aprovechamiento, muy bien pudo haber sido el mismo como resultado de cualquier tipo de currículum. El punto no es si el currículum tuvo o no un efecto, sino que sin una comparación formal y controlada del desempeño de los miembros del grupo "experimental", contra el desempeño de los miembros de algún otro grupo que no experimentó el nuevo currículum, poco es lo que puede decirse acerca de su efecto.

Una distinción importante requiere tomarse en cuenta. No es que el método carezca por completo de valor, sino que *científicamente* carece de valor. En la vida diaria, por supuesto, dependemos de este tipo de evidencia científicamente cuestionable; tenemos que hacerlo. Actuamos, digamos, con base en nuestra experiencia; tenemos la esperanza de que utilizamos nuestra experiencia de forma racional. No se critica el paradigma del pensamiento diario implicado en el diseño 19.1; únicamente que cuando un paradigma de ese tipo se utiliza o se considera científico, entonces comienzan las dificultades. Aun en tareas intelectuales elevadas se utiliza el pensamiento implícito en este diseño. Las observaciones cuidadosas y los análisis brillantes y creativos de Freud sobre el comportamiento neurótico parecen caer dentro de esta categoría. La queja no es en contra de Freud, sino en contra de las suposiciones de que estas conclusiones están "científicamente establecidas".

Diseño 19.2: De un grupo, antes-después (pretest, postest)			
(*a*) Y_a	X	Y_d	(Experimental)
(*b*) Y_a	(X)	Y_d	(No experimental)

El diseño 19.2 representa sólo una pequeña mejoría del diseño 19.1. La característica esencial de esta forma de investigación es que un grupo se compara consigo mismo. Teóricamente, no existe una mejor opción puesto que se controlan todas las variables independientes posibles asociadas con las características de los participantes. El procedimiento sugerido por un diseño de este tipo es el siguiente: se mide un grupo en su variable dependiente, Y, antes de la manipulación experimental, lo cual se llama generalmente *pretest*. Suponga que se miden las actitudes de un grupo de participantes hacia las mujeres; se utiliza una manipulación experimental diseñada para cambiar dichas actitudes. Un experimentador podría exponer al grupo a una opinión experta sobre los derechos de la mujer, por ejemplo. Después de la interposición de esta X, las actitudes de los participantes se miden nuevamente. Se examinan las diferencias de las puntuaciones sobre el cambio de actitud, o $Y_d - Y_a$.

Aparentemente, ésta parecería una buena manera de lograr el propósito experimental. Después de todo, si las diferencias entre las puntuaciones son estadísticamente significativas, ¿esto no indica un cambio en las actitudes? La situación no resulta tan sencilla. Existen otros factores que quizás hayan contribuido al cambio en las puntuaciones; por lo tanto, se confunden los factores. Campbell (1957) ofrece una excelente y detallada discusión de estos factores; aquí únicamente se presenta un resumen sobre ello.

Medición, historia, maduración

Primero está el posible efecto del procedimiento de medición: el hecho de medir a los participantes los cambia. ¿Es posible que las medidas post-X se vieran influenciadas no

mínimo de información científica útil requiere de por lo menos una comparación formal. El ejemplo del currículum requiere, *por lo menos*, una comparación entre el grupo que experimentó el nuevo currículum con otro que no lo haya experimentado. El supuesto efecto del nuevo curriculum, digamos tal y cual aprovechamiento, muy bien pudo haber sido el mismo como resultado de cualquier tipo de currículum. El punto no es si el curriculum tuvo o no un efecto, sino que sin una comparación formal y controlada del desempeño de los miembros del grupo "experimental", contra el desempeño de los miembros de algún otro grupo que no experimentó el nuevo currículum, poco es lo que puede decirse acerca de su efecto.

Una distinción importante requiere tomarse en cuenta. No es que el método carezca por completo de valor, sino que *científicamente* carece de valor. En la vida diaria, por supuesto, dependemos de este tipo de evidencia científicamente cuestionable; tenemos que hacerlo. Actuamos, digamos, con base en nuestra experiencia; tenemos la esperanza de que utilizamos nuestra experiencia de forma racional. No se critica el paradigma del pensamiento diario implicado en el diseño 19.1; únicamente que cuando un paradigma de ese tipo se utiliza o se considera científico, entonces comienzan las dificultades. Aun en tareas intelectuales elevadas se utiliza el pensamiento implícito en este diseño. Las observaciones cuidadosas y los análisis brillantes y creativos de Freud sobre el comportamiento neurótico parecen caer dentro de esta categoría. La queja no es en contra de Freud, sino en contra de las suposiciones de que estas conclusiones están "científicamente establecidas".

Diseño 19.2: De un grupo, antes-después (pretest, postest)

(*a*) Y_a	X	Y_d	(Experimental)
(*b*) Y_a	(X)	Y_d	(No experimental)

El diseño 19.2 representa sólo una pequeña mejoría del diseño 19.1. La característica esencial de esta forma de investigación es que un grupo se compara consigo mismo. Teóricamente, no existe una mejor opción puesto que se controlan todas las variables independientes posibles asociadas con las características de los participantes. El procedimiento sugerido por un diseño de este tipo es el siguiente: se mide un grupo en su variable dependiente, Y, antes de la manipulación experimental, lo cual se llama generalmente *pretest*. Suponga que se miden las actitudes de un grupo de participantes hacia las mujeres; se utiliza una manipulación experimental diseñada para cambiar dichas actitudes. Un experimentador podría exponer al grupo a una opinión experta sobre los derechos de la mujer, por ejemplo. Después de la interposición de esta X, las actitudes de los participantes se miden nuevamente. Se examinan las diferencias de las puntuaciones sobre el cambio de actitud, o $Y_d - Y_a$.

Aparentemente, ésta parecería una buena manera de lograr el propósito experimental. Después de todo, si las diferencias entre las puntuaciones son estadísticamente significativas, ¿esto no indica un cambio en las actitudes? La situación no resulta tan sencilla. Existen otros factores que quizás hayan contribuido al cambio en las puntuaciones; por lo tanto, se confunden los factores. Campbell (1957) ofrece una excelente y detallada discusión de estos factores; aquí únicamente se presenta un resumen sobre ello.

Medición, historia, maduración

Primero está el posible efecto del procedimiento de medición: el hecho de medir a los participantes los cambia. ¿Es posible que las medidas post-X se vieran influenciadas no

sólo por la manipulación de X sino por un incremento en la sensibilización por el pretest? Campbell (1957) llama a dichas medidas *reactivas*, ya que por sí mismas provocan que el sujeto reaccione. Por ejemplo, las actitudes controvertidas parecen ser especialmente susceptibles a dicha sensibilización. Las medidas de aprovechamiento, aunque quizá menos reactivas, también se afectan. Las medidas que involucran a la memoria son susceptibles; si se responde un examen ahora, es más probable que se recuerden las últimas cosas incluidas en el examen. En resumen, los cambios observados pueden deberse a efectos reactivos.

Otras dos fuentes importantes de varianza extraña son la *historia* y la *maduración*. Entre las pruebas Y_a y Y_d pueden ocurrir muchas cosas diferentes a X. A mayor periodo de tiempo, mayor será la posibilidad de que variables extrañas afecten a los participantes y, por lo tanto, a las medidas de Y_d. Esto es lo que Campbell (1957) llama *historia*. Dichas variables o eventos son *específicos* para la situación experimental particular. La *maduración*, por otro lado, cubre eventos que son *generales* —no son específicos de cualquier situación particular, sino que reflejan cambio o crecimiento en el organismo estudiado—. La edad mental se incrementa con el tiempo, un incremento que fácilmente afecta el aprovechamiento, la memoria y las actitudes. La gente puede aprender en cualquier intervalo de tiempo dado, y el aprendizaje puede afectar las medidas de la variable dependiente. Ésta es una de las dificultades exasperantes de la investigación, que perdura por periodos considerables. Mientras más prolongado sea el intervalo de tiempo, mayor será la posibilidad de que fuentes extrañas e indeseables de varianza sistemática influyan en las medidas de la variable dependiente.

El efecto de regresión

Un fenómeno estadístico que ha confundido a los investigadores es el llamado *efecto de regresión*. Las puntuaciones de las pruebas cambian como un hecho de la vida estadística: en el retest, en general, los sujetos tienden a regresar a la media. El efecto de regresión opera a causa de la correlación imperfecta entre las puntuaciones del pretest y del postest. Si $r_{da} = 1.00$, entonces no hay efecto de regresión; si $r_{da} = .00$, entonces el efecto es máximo en el sentido de que la mejor predicción de cualquier puntuación del postest, a partir de la puntuación del pretest, es la media. Con la correlación encontrada en la práctica, el efecto neto es que las puntuaciones más bajas en el pretest tienden a ser altas, y las puntuaciones más altas tienden a ser más bajas en el postest —cuando, de hecho, no ha ocurrido un cambio real en la variable dependiente—. De este modo, si en un estudio se utilizan participantes con bajas puntuaciones, sus puntuaciones en el postest probablemente serán más altas que en el pretest, debido al efecto de regresión. Lo anterior puede engañar al investigador al hacerlo creer que la intervención experimental resultó efectiva, cuando en realidad no fue así. De la misma forma, se puede concluir erróneamente que una variable experimental ha tenido un efecto depresor en los sujetos con altas puntuaciones en el pretest, lo cual no es así necesariamente. Las puntuaciones más altas y más bajas de los dos grupos quizá se deban al efecto de regresión. ¿Cómo funciona esto? Existen muchos factores del azar que influyen en cualquier conjunto de puntuaciones. Dos excelentes referencias sobre la discusión del efecto de regresión son la de Anastasi (1958) y la de Thorndike (1963). Para una presentación más compleja desde el punto de vista estadístico, véase Nesselroade, Stigler y Baltes (1980). En el pretest, algunas puntuaciones altas son mayores de lo que "deberían ser" a causa del azar, y lo mismo sucede con algunas puntuaciones bajas. En el postest es poco probable que se mantengan las puntuaciones altas, ya que los factores que las hicieron altas eran factores del azar —los cuales no están correlacionados en el pretest y postest—. De este modo, el sujeto con una puntuación alta tenderá a bajar en el postest. Un argumento similar se aplica al sujeto con baja puntuación, pero de manera inversa.

Los diseños de investigación deben construirse con el efecto de regresión en mente. No hay manera de controlarlo en el diseño 19.2. Si hubiera un grupo control, entonces se podría "controlar" el efecto de regresión, ya que ambos grupos, el control y el experimental, cuentan con un pretest y un postest. Si la manipulación experimental hubiese tenido un efecto "real", entonces ello debería notarse por encima del efecto de regresión. Es decir, las puntuaciones de ambos grupos, manteniendo igual lo demás, se afectan de la misma manera por la regresión y por otras influencias. Así, si los grupos difieren en el postest, debe ser por la manipulación experimental.

El diseño 19.2 resulta inadecuado, no tanto porque puedan operar variables extrañas y el efecto de regresión (las variables extrañas operan siempre que hay un intervalo de tiempo entre el pretest y el postest), sino *porque no se sabe si éstos han operado, si han afectado las medidas de la variable dependiente.* El diseño no brinda oportunidad alguna para controlar o probar tales posibles influencias.

Diseño 19.3: Simulación de antes-después

	X	Y_a
Y_a		

El título peculiar del diseño 19.3 surge en parte de su propia naturaleza. Como el diseño 19.2, es un diseño antes-después. En lugar de utilizar las mediciones previas y posteriores (o pretest-postest) de un grupo, se emplean como medidas del pretest las medidas de otro grupo, el cual se elige para ser tan similar como sea posible al grupo experimental y, por lo tanto, constituye algo parecido a un grupo control. (La línea entre los dos niveles en el esquema indica grupos separados.) Este diseño satisface la condición de tener un grupo control y, por lo tanto, es un paso más hacia la comparación necesaria en la investigación científica. Por desgracia, los controles son débiles como resultado de la imposibilidad que enfrenta el investigador para saber si los dos grupos eran equivalentes antes de X, la manipulación experimental.

Diseño 19.4: De dos grupos, sin control

(*a*)	X	Y	(Experimental)
	$\sim X$	$\sim Y$	
(*b*)	(X)	Y	(No experimental)
	$(\sim X)$	$\sim Y$	

El diseño 19.4 es común. En (*a*) al grupo experimental se le administra el tratamiento X. El grupo "control", al que se toma o asume como similar al grupo experimental, no recibe X. Las medidas Y se comparan para comprobar el efecto de X. Los grupos o participantes se toman "como son" o pueden ser apareados. La versión no experimental del mismo diseño se clasifica como (*b*). Se observa si un efecto, Y, ocurre en un grupo (línea superior), pero no en otro grupo; o si ocurre en menor grado en el otro grupo (indicado por $\sim Y$ en la línea inferior). Se descubre que el primer grupo experimentó X y el segundo grupo no.

Este diseño tiene una debilidad básica: se *asume* que los dos grupos son iguales respecto a las variables independientes, excepto por X. Algunas veces es posible verificar la igualdad de los grupos de manera general, al compararlos respecto a diferentes variables pertinentes, por ejemplo, edad, sexo, ingresos, inteligencia, habilidad, etcétera. Esto debe hacerse si es posible, pero como Stouffer afirma (1950, p. 522), "con demasiada frecuencia

existe una puerta muy abierta, a través de la cual otras variables no controladas pueden entrar". Puesto que no se utiliza la aleatorización —es decir, los participantes no son asignados aleatoriamente a los grupos—, no es posible suponer que los grupos sean iguales. Ambas versiones del diseño padecen seriamente de falta de control de las variables independientes por la falta de aleatorización.

Criterios del diseño de investigación

Después de examinar algunas de las principales debilidades de los diseños de investigación inadecuados, ahora es un buen momento para discutir lo que puede llamarse *criterios* del diseño de investigación. Junto con los criterios se enunciarán ciertos principios para guiar a los investigadores. Por último, los criterios y principios se relacionarán con las nociones de validez interna y externa de Campbell (1957), las cuales en cierto sentido expresan los criterios de otra forma.

¿Responder preguntas de investigación?

El criterio principal de un diseño de investigación puede expresarse en una pregunta: *¿el diseño responde a la pregunta de investigación?* O *¿el diseño prueba adecuadamente las hipótesis?* Quizá la debilidad más seria de los diseños, con frecuencia propuesta por los neófitos, es que no son capaces de responder adecuadamente las preguntas de investigación. Un ejemplo común de esta falta de congruencia entre las preguntas de investigación y las hipótesis, por un lado, y el diseño de investigación, por el otro, es el apareamiento de los participantes por razones que son irrelevantes a la investigación, y luego el uso de un grupo experimental del tipo de diseño con grupo control. Por ejemplo, los estudiantes a menudo suponen que, debido a que aparean a los sujetos con respecto a inteligencia y género, sus grupos experimentales son iguales. Ellos han escuchado que se requiere aparear a los participantes como "control" y que se necesita un grupo experimental y un grupo control. Sin embargo, frecuentemente las variables apareadas resultan irrelevantes para los propósitos de la investigación. Es decir, si no existe relación entre, digamos, el género y la variable dependiente, el apareamiento por género es irrelevante.

Otro ejemplo de esta debilidad es el caso donde se necesitan tres o cuatro grupos experimentales. Por ejemplo, con tres grupos experimentales y un grupo control, o cuatro grupos con diferentes cantidades o aspectos de X, se requiere el tratamiento experimental. Sin embargo, el investigador usa sólo dos porque ha escuchado que un grupo experimental y un grupo control son necesarios y deseables.

El ejemplo que se presentó en el capítulo 18, referente a la comprobación de una hipótesis de interacción realizando dos experimentos separados, es otro ejemplo. La hipótesis a prueba era que la discriminación en las admisiones a la universidad es una función de género y del nivel de habilidad; que se excluye a las mujeres con baja habilidad (en contraste con los hombres de baja habilidad). Ésta es una hipótesis de interacción y probablemente requiera de un diseño de tipo factorial. Establecer dos experimentos, uno para los aplicantes con alta habilidad y otro para los aplicantes con baja habilidad, constituye un procedimiento pobre porque dicho diseño, como se mostró anteriormente, no prueba en definitiva la hipótesis planteada. De la misma manera, aparear a los participantes respecto a su habilidad y después establecer un diseño de dos grupos, perdería por completo la pregunta de investigación. Tales consideraciones conducen a un precepto general y aparentemente obvio:

Diseñar la investigación para responder preguntas de investigación.

Control de variables independientes extrañas

El segundo criterio es el *control*, que se refiere al control de variables independientes: las variables independientes del estudio de investigación y las variables independientes extrañas. Las variables independientes extrañas son, por supuesto, variables que pueden influir en la variable dependiente; pero que no son parte del estudio. Dichas variables se confunden con la variable independiente bajo estudio. En el estudio sobre admisiones del capítulo 18, por ejemplo, la ubicación geográfica (de las universidades) quizá sea una variable extraña potencialmente influyente, que opaque los resultados del estudio. Es decir, si las universidades del este rechazan más mujeres que las universidades del oeste, entonces la localización geográfica es una fuente de varianza extraña en las medidas de admisión, que debe ser controlada de alguna manera. El criterio se refiere también al control de las variables del estudio. Ya que este problema se ha discutido y continuará en análisis, no es necesario decir más aquí. Pero la pregunta debe plantearse: *¿este diseño controla adecuadamente las variables independientes?*

La mejor forma de responder satisfactoriamente esta pregunta se expresa en el siguiente principio:

> *Aleatorizar siempre que sea posible: seleccionar aleatoriamente a los participantes; asignar aleatoriamente a los participantes a los grupos; asignar aleatoriamente los tratamientos experimentales a los grupos.*

Mientras que quizá no sea posible seleccionar aleatoriamente a los participantes, puede ser posible asignarlos aleatoriamente a los grupos, "igualando" así los grupos en el sentido estadístico analizado en capítulos previos. Si tal asignación aleatoria de los participantes a los grupos no es factible, entonces debe realizarse un gran esfuerzo para asignar aleatoriamente los tratamientos experimentales a los grupos experimentales. Y, si los tratamientos experimentales se administran en diferentes momentos con diferentes experimentadores, entonces los momentos y los experimentadores deben asignarse de forma aleatoria.

El principio que vuelve pertinente la aleatorización es complejo y difícil de aplicar:

> *Controlar las variables independientes para que las fuentes extrañas e indeseables de varianza sistemática tengan la mínima oportunidad de operar.*

Como se aprendió antes (capítulo 8), en teoría la aleatorización satisface este principio. Cuando se prueba la validez empírica de una proposición: si p entonces q, se manipula p y se observa que q covaría con la manipulación de p. ¿Pero qué tanta confianza se puede tener en que la proposición si p entonces q sea realmente "verdadera"? La confianza está directamente relacionada con qué tan completos y adecuados son los controles. Si se utiliza un diseño similar a los diseños 19.1 a 19.4, no se puede tener demasiada confianza en la validez empírica de la proposición si p entonces q, debido a que el control de variables independientes extrañas es débil o inexistente. Puesto que dicho control no es siempre posible en gran parte de la investigación psicológica, sociológica y educativa, entonces, ¿hay que abandonar la investigación por completo? En lo absoluto. Sin embargo, es necesario estar consciente de las debilidades del diseño intrínsecamente pobre.

Posibilidad de generalización

El tercer criterio, la *generalización*, es independiente de otros criterios pues es diferente tipo. Éste es un punto importante que pronto quedará claro. Tan sólo significa: *¿es posible generalizar los resultados de un estudio a otros participantes, otros grupos y otras*

condiciones? Quizá la pregunta se plantee mejor así: *¿qué tanto* pueden generalizarse los resultados del estudio? Quizás ésta sea la pregunta más compleja y difícil que pueda hacerse respectoa los datos de investigación, ya que no sólo toca cuestiones técnicas (como el muestreo y el diseño de investigación), sino también problemas más amplios de la investigación básica y aplicada. En la investigación básica, por ejemplo, la posibilidad de generalización no es la primera consideración, pues el interés central son las relaciones entre variables y por qué tales variables se relacionan como lo hacen. Lo anterior enfatiza los aspectos internos en vez de los aspectos externos del estudio. Estos estudios frecuentemente se diseñan para examinar cuestiones teóricas tales como la motivación o el aprendizaje. La meta de la investigación básica consiste en aportar información y conocimiento a un campo de estudio pero, en general, sin un propósito práctico específico. Sus resultados son generalizables; aunque no en el mismo terreno que los resultados encontrados en estudios de investigación aplicada, en la cual, por otro lado, el interés central obliga a preocuparse más por la generalización, puesto que en efecto se desea aplicar los resultados a otras personas y a otras situaciones. Los estudios de investigación aplicada por lo común se fundamentan en estudios de investigación básica. Con el uso de información encontrada en un estudio de investi-gación básica, los estudios de investigación aplicada utilizan dichos hallazgos para determinar si pueden resolver un problema práctico. Por ejemplo, considere el trabajo de **B. F.** Skinner; sus primeras investigaciones son generalmente consideradas como investigación básica. Fue a partir de su investigación que los programas de reforzamiento fueron establecidos. Sin embargo, más adelante Skinner y otros (Skinner, 1968; Garfinkle, Kline y Stancer, 1973) aplicaron los programas de reforzamiento a problemas militares, educativos y de comportamiento. Quienes realizan investigación sobre la modificación del comportamiento están aplicando muchas de las teorías e ideas probadas y establecidas por B. F. Skinner. Si el lector pondera los siguientes dos ejemplos de investigación básica y aplicada, entonces podrá acercarse a esta distinción.

En el capítulo 14 se examinó un estudio de Johnson (1994) respecto al tipo de violación, admisibilidad de información y percepción de las víctimas de violación. Esta es claramente investigación básica: el interés central fueron las relaciones entre el tipo de violación, admisibilidad de información y percepción. A pesar de que nadie sería tan insensato para afirmar que a Johnson no le preocupaba el tipo de violación, la admisibilidad de información y la percepción, en general, el énfasis recayó en las relaciones entre las variables del estudio. Contraste este estudio con el esfuerzo de Walster *et al.* (1970) para determinar si las universidades discriminan en contra de las mujeres. Naturalmente, Walster y sus colegas fueron exigentes respecto a los aspectos internos de su estudio; pero ellos por fuerza debían tener otro interés: ¿se practica la discriminación entre las universidades en general? Su estudio es claramente investigación aplicada, aunque no puede decirse que había ausencia de interés de realizar investigación básica. Las consideraciones de la siguiente sección ayudan a explicar la posibilidad de generalización.

Validez interna y externa

Dos criterios generales del diseño de investigación se han discutido con profundidad por Campbell (1957) y por Campbell y Stanley (1963). Estos conceptos constituyen una de las contribuciones más significativas, importantes e informativas a la metodología de la investigación durante las pasadas tres o cuatro décadas.

La *validez interna* plantea la pregunta: ¿La manipulación experimental, *X*, realmente causó una diferencia significativa? Los tres criterios del capítulo 18 en realidad son aspectos de la validez interna. De hecho, cualquier cuestión que afecte los *controles* de un diseño

se convierte en un problema de validez interna. Si un diseño es tal que sólo es posible tener poca o ninguna confianza en las relaciones, como se muestra por las diferencias significativas entre grupos experimentales, entonces se trata de un problema de validez interna.

Con anterioridad en este capítulo se presentaron cuatro amenazas posibles a la validez interna. Algunos autores de libros de texto se han referido a ello como "explicaciones alternativas" (véase Dane, 1990) o "hipótesis rivales" (véase Graziano y Raulin, 1993). Éstas fueron enlistadas como medición, historia, maduración y regresión estadística. Campbell y Stanley (1963) enlistan también otras cuatro amenazas: instrumentación, selección, abandono y la interacción entre algunas de las amenazas mencionadas (un total de ocho).

La instrumentación es un problema del dispositivo utilizado para medir los cambios de la variable dependiente a través del tiempo. Esto es especialmente verdadero en estudios que usan observadores humanos. Los observadores humanos o jueces quizá se vean afectados por eventos previos o por fatiga. Los observadores pueden volverse más eficientes a través del tiempo, de tal manera que las últimas mediciones sean más precisas que las primeras. Por otro lado, los observadores humanos con fatiga se volverían menos precisos en los últimos ensayos que en los primeros; cuando así sucede, los valores de la variable dependiente cambiarán, y dicho cambio no se deberá sólo a la manipulación de la variable independiente.

Con el término selección, Campbell y Stanley (1963) se refieren al tipo de participantes que el experimentador selecciona para el estudio, lo cual ocurre así cuando el investigador no es precavido en estudios que no utilizan selección o asignación aleatorias. El investigador pudo haber seleccionado participantes en cada grupo que fueran muy diferentes en algunas características y, de esta manera, encontrar una diferencia en la variable dependiente. Es importante que el investigador tenga igualados los grupos previamente a la administración del tratamiento. Si los grupos son iguales antes del tratamiento, la lógica indica que si son diferentes después del tratamiento, entonces fue el tratamiento (variable independiente) lo que causó las diferencias y no otra cosa. Sin embargo, si los grupos son diferentes al inicio y diferentes después del tratamiento, es muy difícil afirmar que la diferencia se debió al tratamiento. Más adelante, cuando se discutan los diseños cuasiexperimentales, se verá cómo se puede fortalecer la situación.

El abandono o la mortandad experimental se refiere al retiro de los participantes. Si demasiados participantes en una condición de tratamiento abandonan el estudio, el desequilibrio se vuelve una posible razón del cambio en la variable dependiente. El abandono también incluye la salida de los participantes con ciertas características.

Cualquiera de estas siete amenazas a la validez interna también pueden interactuar entre sí. La selección puede interactuar con la maduración. La amenaza es especialmente posible cuando se utilizan participantes voluntarios. Si el investigador está comparando dos grupos —un grupo formado por voluntarios (autoseleccionados) y el otro grupo, por no voluntarios— la diferencia en el desempeño de ambos en la variable dependiente quizá se deba al hecho de que los voluntarios estén más motivados. Los investigadores estudiantes algunas veces utilizan sujetos voluntarios o miembros de su propia familia o círculo social como participantes. Puede haber un problema de validez interna si se ubica a los voluntarios en un grupo de tratamiento y si sus amigos son colocados en otro.

Un criterio difícil de satisfacer —*la validez externa*— es la *representatividad* o *posibilidad de generalización*. Cuando se completa un estudio y se encuentra una relación, ¿a qué población podría generalizarse? ¿Puede decirse que *A* se relaciona con *B* en todos los alumnos? ¿En todos los alumnos de octavo grado? ¿En todos los alumnos de octavo grado en este sistema escolar? O, ¿en todos los alumnos de octavo grado únicamente de esta

escuela? ¿O los hallazgos deben limitarse a los alumnos de octavo grado con quienes se trabajó? Siempre *deben preguntarse y responderse* estas importantes preguntas científicas.

No sólo debe cuestionarse la generalización de la muestra, sino que también es necesario plantear preguntas acerca de la representatividad ecológica y de las variables de los estudios. Si cambia el escenario social donde se condujo el experimento, ¿se mantendrá aún la relación entre *A* y *B*? ¿Estarán *A* y *B* relacionadas si se replica el estudio en una escuela de menor clase social? ¿En una escuela occidental? ¿En una escuela del sur? Las anteriores son preguntas sobre la *representatividad ecológica*.

La *representatividad de la variable* es un término más sutil. Una pregunta que no se plantea frecuentemente, pero que debe hacerse, es: ¿las variables de este estudio son representativas? Cuando un investigador trabaja con variables psicológicas y sociológicas, se asume que las variables son "constantes". Si el investigador encuentra una diferencia en aprovechamiento entre niños y niñas, se asume que el género, como variable, es "constante".

En el caso de variables como aprovechamiento, agresión, aptitud y ansiedad, ¿el investigador puede suponer que la "agresión" de los participantes suburbanos es la misma "agresión" que se encontraría en los barrios bajos citadinos? ¿La variable es igual en los suburbios europeos? La representatividad de la "ansiedad" es más difícil de determinar. Cuando se habla de "ansiedad", ¿a qué tipo de ansiedad se refiere? ¿Todos los tipos de ansiedad son iguales? Si la ansiedad se manipula en una situación por medio de instrucciones verbales y en otra situación por medio de un choque eléctrico, ¿las dos ansiedades inducidas son iguales? Si la ansiedad es manipulada por, digamos, la instrucción experimental, ¿ésta es la misma ansiedad que la medida por medio de una escala de ansiedad? La representatividad de las variables es, entonces, otro aspecto del gran problema de la validez externa y, por lo tanto, de la generalización.

A menos que se tomen precauciones especiales y que se realicen esfuerzos considerables, los resultados de investigación con frecuencia no son representativos y, por lo tanto, no son generalizables. Campbell y Stanley (1963) afirman que la validez interna es una condición indispensable del diseño de investigación, pero que el diseño ideal debe ser fuerte tanto en la validez interna como en la externa, aun cuando éstas sean frecuentemente contradictorias. En estos capítulos el principal énfasis recae en la validez interna, con un ojo vigilante sobre la validez externa.

Campbell y Stanley (1963) presentan cuatro amenazas a la validez externa. Son los efectos reactivos o de interacción de la prueba, los efectos de interacción de los sesgos de selección y la variable independiente, los efectos reactivos de los arreglos experimentales y la interferencia de tratamiento múltiple.

El efecto reactivo o de interacción de la prueba se refiere al uso de un pretest antes de la administración del tratamiento. Aplicar un pretest quizá disminuya o incremente la sensibilidad del participante a la variable independiente; esto haría que los resultados de la población evaluada en el pretest no sean representativos del efecto del tratamiento para la población que no fue evaluada con anterioridad. La probabilidad de una interacción entre el tratamiento y el pretest parece haber sido señalada en primera instancia por Solomon (1949).

El efecto de interacción del sesgo en la selección y la variable independiente indica que la selección de los participantes puede muy bien afectar la generalización de los resultados. Un investigador que utiliza sólo participantes de la población de sujetos de una universidad en particular, que generalmente consiste de estudiantes de primero y de segundo año, encontrará difícil generalizar los resultados del estudio a otros alumnos de la universidad o de otras universidades.

La mera participación en un estudio de investigación puede constituir un problema en términos de la validez externa. La presencia de observadores, la instrumentación o el

ambiente de laboratorio pueden tener un efecto sobre el participante, lo que no ocurriría si el participante estuviera en un escenario natural. El hecho de participar en un estudio experimental puede alterar la conducta normal del sujeto. Si el experimentador es hombre o mujer, afroamericano o blanco también puede tener un efecto.

Si los participantes son expuestos a más de una condición de tratamiento, el desempeño en ensayos posteriores se ve afectado por el desempeño en los primeros ensayos. Por lo tanto, los resultados sólo pueden generalizarse a personas que han tenido múltiples exposiciones, presentadas en el mismo orden.

El enfoque negativo de este capítulo se hizo con la creencia de que una exposición sobre procedimientos pobres, pero comúnmente utilizados y *aceptados*, junto con una discusión sobre sus mayores debilidades, proporcionarían un buen punto de inicio para el estudio del diseño de investigación. Otros diseños inadecuados son posibles; aunque todos ellos son inadecuados únicamente en sus principios estructurales. Dicho punto debe enfatizarse, ya que en el capítulo 20 se observará que una estructura de diseño perfecta puede ser utilizada pobremente. Por lo tanto, es necesario aprender y entender las dos fuentes de debilidad en la investigación: los diseños intrínsecamente pobres y los diseños intrínsecamente buenos pero pobremente utilizados.

RESUMEN DEL CAPÍTULO

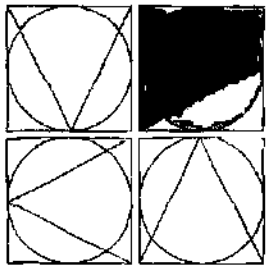
1. El estudio de diseños inadecuados ayuda al investigador a diseñar mejores estudios al saber qué dificultades evitar.
2. Los diseños no experimentales son aquellos con variables independientes no manipuladas y con ausencia de asignación o selección aleatorias.
3. Los diseños inadecuados incluyen el diseño de "estudio de caso de un disparo", el diseño pretest-postest, el diseño pretest-postest simulado y el diseño de dos grupos sin control.
4. Los diseños inadecuados se analizan en términos de su validez interna.
5. La validez interna implica qué tanto el experimentador puede establecer el efecto de la variable independiente sobre la variable dependiente. A mayor confianza del investigador respecto a la variable independiente manipulada, más fuerte será la validez interna.
6. Los estudios no experimentales son más débiles en cuanto a la validez interna que los estudios experimentales.
7. Existen ocho clases básicas de variables extrañas, las cuales, si no son controladas, pueden confundirse con la variable independiente. Las ocho clases básicas se denominan amenazas a la validez interna.
8. Las amenazas a la validez interna, según Campbell, se enumeran como sigue:
   - Historia
   - Maduración
   - Prueba o medición
   - Instrumentación
   - Regresión estadística
   - Selección
   - Mortalidad experimental o abandono
   - Interacción selección-maduración
9. La validez externa implica qué tan fuerte es la afirmación que el experimentador puede hacer respecto a la generalización de los resultados del estudio.

10. Campbell y Stanley señalan cuatro fuentes posibles de amenaza a la validez externa:
    - Efecto reactivo o de interacción de la prueba
    - Efectos de interacción de los sesgos de selección y la variable independiente
    - Efectos reactivos de los arreglos experimentales
    - Interferencia de tratamiento múltiple

SUGERENCIAS DE ESTUDIO

1. Suponga que una universidad de arte decide iniciar un nuevo currículum para todos los estudiantes de pregrado. Se pide al profesorado formar un grupo de investigación para estudiar la efectividad del programa durante dos años. Con el objetivo de tener un grupo con el cual comparar al grupo del nuevo currículum, el grupo de investigación solicita que el programa actual se continúe por dos años y que se permita a los estudiantes elegir el programa actual o el nuevo. El grupo de investigación considera que así tendrán un grupo experimental y un grupo control.

   Analice críticamente la propuesta del grupo de investigación. ¿Qué tanta confianza tendría usted en los hallazgos al final de los dos años? Mencione las razones de su reacción positiva o negativa hacia la propuesta.
2. Imagine que usted es profesor de una escuela de posgrado y se le pide juzgar el valor de una tesis doctoral propuesta. La estudiante de doctorado es una jefa escolar que está instituyendo un nuevo tipo de administración dentro de su sistema escolar. Ella planea estudiar los efectos de la nueva administración durante un periodo de tres años y, después, escribir la tesis. Ella dice que no estudiará ninguna otra situación escolar durante el periodo para no sesgar los resultados. Discuta la propuesta y, al hacerlo, plantéese la pregunta: ¿la propuesta es adecuada para un trabajo doctoral?
3. En su opinión, ¿debe basarse estrictamente toda investigación en el criterio de generalización? Explique por qué sí o por qué no. ¿Qué campo puede ser que tenga más investigación básica: psicología o educación? ¿Por qué? ¿Qué implicaciones tienen sus conclusiones para la generalización?
4. ¿Qué tiene que ver la replicación de investigación con la generalización? Explique. Si fuese posible, ¿debería replicarse toda investigación? Explique por qué sí o por qué no. ¿Qué tiene que ver la replicación con la validez interna y externa?



CAPÍTULO 20

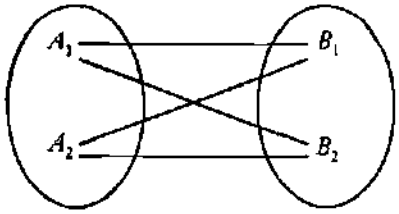
DISEÑOS GENERALES DE INVESTIGACIÓN

El diseño constituye una disciplina de datos. El propósito implícito de todo diseño de investigación consiste en imponer restricciones controladas sobre observaciones de fenómenos naturales. El diseño de investigación, en efecto, le dice al investigador: Haga esto y aquello; no haga esto ni aquello; tenga cuidado con esto; ignore aquello; etcétera. Es el proyecto del arquitecto e ingeniero de investigación. Si el diseño está concebido de forma estructuralmente pobre, el producto final será defectuoso. Si al menos está bien concebido desde el punto de vista estructural, el producto final tiene una mayor probabilidad de alcanzar atención científica seria. En este capítulo, la principal preocupación son distintos diseños básicos de investigación "buenos". También se analizan ciertos fundamentos conceptuales de investigación y algunos problemas relacionados con el diseño; por ejemplo, la lógica de los grupos control y los pros y los contras del apareamiento.

Fundamentos conceptuales del diseño de investigación

Los fundamentos conceptuales para entender el diseño de investigación se establecieron en los capítulos 4 y 5, donde se definieron y analizaron los conjuntos y las relaciones. Recuerde que una *relación* es un conjunto de pares ordenados y también que un *producto cartesiano* son todos los pares ordenados posibles de dos conjuntos. Una *partición* divide un conjunto universal U en subcojuntos que están *separados* y son *exhaustivos*. Una *partición cruzada* es una partición nueva que surge de partir sucesivamente U formando todos los subconjuntos de la forma $A \cap B$. Tales definiciones se explicaron en los capítulos 5 y 6. Ahora se aplicarán al diseño y a las ideas de análisis.

Tome dos conjuntos, A y B, divididos en A_1 y A_2, B_1 y B_2. El producto cartesiano de los dos conjuntos es:



Los pares ordenados, entonces, son: A_1B_1, A_1B_2, A_2B_1, A_2B_2. Puesto que hay un conjunto de pares ordenados, esto es, una relación; también es una partición cruzada. El lector debe revisar las figuras 4.7 y 4.8 del capítulo 4 para ayudarse a aclarar estas ideas y para conocer la aplicación del producto cartesiano y las ideas de relación en el diseño de investigación. Por ejemplo, A_1 y A_2 pueden ser dos aspectos de cualquier variable independiente: experimental-control, dos métodos, hombre y mujer, etcétera.

Un *diseño* es algún subconjunto del producto cartesiano de las variables independientes y la variable dependiente. Es posible aparear cada medida de la variable dependiente, a la cual se le llama Y en este análisis, con algún aspecto o partición de una variable independiente. Los casos más simples posibles ocurren con una variable independiente y una variable dependiente. En el capítulo 10, una variable independiente, A, y una variable dependiente, B, se dividieron en $[A_1, A_2]$ y $[B_1, B_2]$, y después se realizó una partición cruzada para formar la ahora familiar tabulación cruzada de 2×2, con frecuencias o porcentajes en las casillas. Sin embargo, el interés está en particiones cruzadas similares de A y B; pero con medidas continuas en las casillas.

Tome solamente a A, utilizando un diseño de análisis de varianza de un factor. Suponga que se tienen tres tratamientos experimentales, A_1, A_2 y A_3 y, para simplificar, dos puntuaciones Y en cada casilla, lo cual se presenta a la izquierda de la figura 20.1, denominada (*a*). Digamos que seis participantes han sido asignados aleatoriamente a tres tratamientos y que las puntuaciones de los seis individuos después de los tratamientos *experimentales* son las que aparecen en la figura.

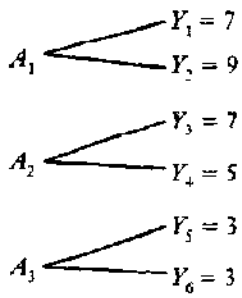
La parte derecha de la figura 20.1, designada como (*b*), muestra la misma idea pero como pares ordenados o en forma de relación. Los pares ordenados son A_1Y_1, A_1Y_2, A_2Y_3, ..., A_3Y_6. Esto no es, por supuesto, un producto cartesiano, el cual aparearía a A_1 con todas las puntuaciones Y, a A_2 con todas las puntuaciones Y y A_3 con todas las puntuaciones Y, dando un total de $3 \times 6 = 18$ pares. De manera más precisa, la figura 20.1(*b*) es un subconjunto del producto cartesiano $A \times B$. Los diseños de investigación son subconjuntos de $A \times B$, y el diseño y el problema de investigación definen o especifican cómo se establecen los subconjuntos. Los subconjuntos del diseño de la figura 20.1 están presumiblemente dictados por el problema de investigación.

FIGURA 20.1

(*a*)

A_1	A_2	A_3
7	7	3
9	5	3

(*b*)



Cuando hay más de una variable independiente, el problema se vuelve más complejo. Tome dos variables independientes, *A* y *B*, divididas en $[A_1, A_2]$ y $[B_1, B_2]$. El lector no debe confundir esto con el paradigma de frecuencias *AB* previo, en el cual *A* era la variable independiente y *B* la variable dependiente.

Ahora deben tenerse tríos ordenados (o dos conjuntos de pares ordenados): *ABY*. Analice la figura 20.2; en el costado izquierdo de la figura, denominada (*a*) se presenta el diseño de 2 × 2 del análisis de varianza factorial y el ejemplo utilizado en el capítulo 14 (véase la figura 14.2, y las tablas 14.3 y 14.4), con las medidas de la variable dependiente *Y*, insertadas en las casillas; es decir, ocho participantes fueron asignados aleatoriamente a las cuatro casillas. Sus puntuaciones después del experimento, son $Y_1, Y_2, \ldots, Y_8$. El costado derecho de la figura, denominado (*b*), muestra los tríos ordenados, *ABY*, como un árbol. Obviamente éstos son subconjuntos de $A \times B \times Y$, y son relaciones. El mismo razonamiento es extensible a diseños más grandes y más complejos, como un factorial de 2 × 2 × 3 (*ABCY*) o uno de 4 × 3 × 2 × 2 (*ABCDY*). (En dichas designaciones generalmente se omite *Y* debido a que está implícita.) Otros tipos de diseños se conceptualizan de manera similar, aunque su descripción por medio de árboles puede resultar laboriosa.

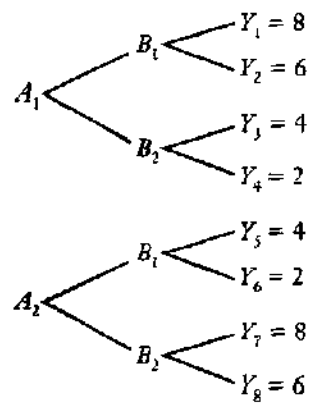
En resumen, un diseño de investigación es algún subconjunto del producto cartesiano de las variables independiente y dependiente. Con sólo una variable independiente, se

FIGURA 20.2

(*a*)

	B_1	B_2
A_1	$Y_1 = 8$ $Y_2 = 6$	$Y_3 = 4$ $Y_4 = 2$
A_2	$Y_5 = 4$ $Y_6 = 2$	$Y_7 = 8$ $Y_8 = 6$

(*b*)



divide la única variable; con más de una variable independiente, las variables independientes se convierten en particiones cruzadas. Con tres o más variables independientes la conceptualización es la misma; sólo difieren las dimensiones, por ejemplo, $A \times B \times C$ y $A \times B \times C \times D$, y sus particiones cruzadas. Siempre que sea posible, es deseable tener diseños "completos" —un diseño completo es una partición cruzada de las variables independientes— y observar las dos condiciones básicas de separación y exhaustividad; es decir, los diseños no deben tener un caso (la puntuación de un participante) en más de una casilla de una partición o de una partición cruzada; y deben utilizarse todos los casos. Además, el mínimo básico de cualquier diseño es, por lo menos, una partición de la variable independiente en dos subconjuntos, por ejemplo, partir A en A_1 y A_2. Existen también diseños "incompletos", aunque en este libro se enfatizan los diseños "completos". Véase Kirk (1995) para estudiar de manera más profunda los diseños incompletos.

El término "diseños generales" establece que los diseños incluidos en el capítulo se simbolizan o expresan en su forma más general y abstracta. Cuando se da una X sencilla (que representa una variable independiente), debe tomarse como un indicador de más de una X —es decir, la X se divide en dos o más grupos experimentales—. Por ejemplo, el diseño 20.1, que se estudiará en breve, tiene X y $\sim X$, que quiere decir que existen grupos control y experimental y, por lo tanto, es una partición de X. Pero X puede dividirse en varias categorías de X, quizás cambiando el diseño, de uno simple con una variable a, digamos, un diseño factorial. No obstante, la simbología básica asociada con el diseño 20.1 permanece igual. Tales complejidades se aclararán en éste y posteriores capítulos.

Una nota preliminar: diseños experimentales y análisis de varianza

Antes de examinar los diseños de este capítulo, es necesario aclarar uno o dos puntos confusos y potencialmente controversiales que por lo común no se consideran en la literatura. La mayoría de los diseños que aquí se estudian son experimentales. Como generalmente se piensa, la lógica de los diseños de investigación está basada en condiciones e ideas experimentales; también están íntimamente ligados a los paradigmas del análisis de varianza. Esto, por supuesto, no es accidental. Las concepciones modernas del diseño, en especial los diseños factoriales, nacieron cuando se inventó el análisis de varianza. Aunque no existe una ley dura que diga que el análisis de varianza sea aplicable únicamente a situaciones experimentales —de hecho, ha sido utilizado muchas veces en investigación no experimental—, por lo común es verdad que resulta más apropiado para los datos de experimentos. Esto es especialmente cierto para los diseños factoriales, donde hay igual número de casos en las casillas del paradigma del diseño, y donde los participantes son asignados aleatoriamente a las condiciones experimentales (o casillas).

Cuando no es posible asignar aleatoriamente a los participantes, y cuando, por una razón u otra, existe un número desigual de casos en las casillas de un diseño factorial, el uso del análisis de varianza se vuelve cuestionable e incluso inapropiado. También puede ser torpe y poco elegante. Lo anterior es así porque el uso del análisis de varianza supone que las correlaciones entre las variables independientes de un diseño factorial son iguales a cero. La asignación aleatoria hace que se mantenga este supuesto, ya que dicha asignación presumiblemente divide las fuentes de varianza de forma igualitaria entre las casillas. Sin embargo, la asignación aleatoria sólo puede lograrse en experimentos. En la investigación no experimental, las variables independientes son características más o menos estables de los participantes (por ejemplo, inteligencia, sexo, clase social y otras similares), que por lo común están correlacionadas sistemáticamente. Considere dos variables indepen-

dientes manipuladas, digamos, reforzamiento y ansiedad. Puesto que los participantes con cantidades variables de características correlacionadas con tales variables están distribuidos aleatoriamente en las casillas, se supone que las correlaciones entre aspectos de reforzamiento y de ansiedad son iguales a cero. Si, por el otro lado, las dos variables independientes son inteligencia y clase social, ambas generalmente no manipuladas y correlacionadas, no se cumple el supuesto de correlación cero entre ellas, necesario para el análisis de varianza. Debe utilizarse algún método de análisis que justifique la correlación entre ellas. Se verá más adelante en el libro que está disponible un método de dicho tipo: la regresión múltiple.

Hasta ahora no se ha alcanzado un estado de madurez de investigación para apreciar la profunda diferencia entre las dos situaciones. Por ahora, sin embargo, acepte la diferencia y la afirmación de que el análisis de varianza es básicamente una concepción y una forma de análisis experimentales. Estrictamente hablando, si las variables independientes son no experimentales, entonces el análisis de varianza no es el tipo de análisis apropiado. No obstante, existen excepciones a esta afirmación; por ejemplo, si una variable independiente es experimental y la otra es no experimental, el análisis de varianza es apropiado. Además, en el análisis de varianza de un factor, ya que sólo hay una variable independiente, el análisis de varianza puede utilizarse con una variable independiente no experimental, aunque quizás el análisis de regresión sería más apropiado. En el número 3 de las sugerencias para estudio se cita un uso interesante del análisis de varianza con datos no experimentales.

De manera similar, si por alguna razón el número de casos en las casillas no es igual (y es desproporcionado), entonces habrá correlación entre las variables independientes y no es sostenible el supuesto de la correlación cero. Esta abstracta y abstrusa digresión del tema principal del diseño puede parecer un poco confusa en este punto del estudio; el problema involucrado deberá quedar claro después de estudiar la investigación experimental y la no experimental y, posteriormente en el libro, el fascinante y poderoso enfoque conocido como regresión múltiple.

Los diseños

En lo que resta de este capítulo se analizan varios diseños básicos de investigación. Recuerde que un diseño es un plan, un proyecto para conceptualizar la estructura de las relaciones entre las variables de un estudio de investigación. Un diseño no sólo dispone las relaciones del estudio, también implica cómo se controla la situación de investigación y cómo se analizarán los datos. Un diseño, en el sentido utilizado en este capítulo, constituye el armazón de la investigación, el cual se recubre con las variables y relaciones de la misma. Los esquemas presentados en los diseños 20.1 a 20.8 representan la estructura simple y abstracta de la investigación. Algunas veces las tablas analíticas, como la figura 20.2 (a la izquierda) y las figuras del capítulo 18 (por ejemplo, las figuras 18.2, 18.3 y 18.5) se llaman *diseños*. Mientras que el llamarles diseños no causa mucho daño, estrictamente hablando son paradigmas analíticos. Sin embargo, para no ser demasiado exigentes, a ambos tipos de representaciones se les llamará "diseños".

Diseño 20.1: Grupo experimental-grupo control: participantes aleatorizados

[*A*]	*X*	*Y*	(Experimental)
	~X	*Y*	(Control)

El diseño 20.1, con dos grupos como en la tabla anterior, y sus variantes con más de dos grupos, probablemente son los "mejores" diseños para muchos propósitos experimentales en la investigación del comportamiento. Campbell y Stanley (1963) llaman a este diseño el *diseño de grupo control sólo con postest*, mientras que Isaac y Michael (1987) se refieren a él como *diseño de grupo control aleatorizado sólo con postest*. La [*A*] que aparece antes del paradigma indica que los participantes son asignados aleatoriamente al grupo experimental (línea superior) y al grupo control (línea inferior). Dicha aleatorización elimina las objeciones al diseño 19.4 mencionado en el capítulo 19. En teoría todas las variables independientes posibles están controladas; por supuesto que en la práctica puede no ser así. Si se incluyen suficientes participantes en el experimento para darle a la aleatorización una oportunidad de "operar", entonces se tiene un control fuerte y se satisfacen bastante bien las demandas de validez interna. Este diseño controla los efectos de la historia, la maduración y el pretest, pero no mide tales efectos.

Si se extiende a más de dos grupos y si es capaz de responder las preguntas de investigación planteadas, el diseño 20.1 tiene diversas ventajas: 1) tiene el mejor sistema de control teórico integrado que cualquier otro diseño, con una o dos excepciones posibles en casos especiales; 2) resulta flexible, es teóricamente capaz de extenderse a cualquier número de grupos con cualquier número de variables; 3) si se extiende a más de una variable, puede probar varias hipótesis al mismo tiempo; y 4) es elegante estadística y estructuralmente.

Antes de estudiar otros diseños, es necesario examinar el concepto de grupo control, una de las invenciones creativas de los pasados cien años, y ciertas extensiones del diseño 20.1.

La noción del grupo control y las extensiones del diseño 20.1

Evidentemente el término *control* y la expresión "grupo control" no aparecían en la literatura científica a finales del siglo XIX, lo cual está documentado por Boring (1954). Sin embargo, la noción de experimentación controlada es más antigua. Boring afirma que Pascal lo utilizó tan temprano como 1648. Solomon (1949) buscó en la literatura psicológica y no pudo encontrar un solo caso del uso de grupo control antes de 1901. Quizá la noción de grupo control fue utilizada en otros campos, aunque es dudoso que la idea estuviese bien desarrollada. Solomon (p. 175) también señala que el estudio sobre actitudes de Peterson y Thurstone de 1933 fue el primer intento serio por emplear grupos control en la evaluación de los efectos de procedimientos educativos. No es posible encontrar la expresión "grupo control" en la famosa decimoprimera edición (1911) de la *Encyclopedia Britannica*, aunque sí se analiza el método experimental. Solomon también afirma que el diseño de grupo control aparentemente tuvo que esperar desarrollos estadísticos y el avance de la sofisticación estadística entre los psicólogos.

Quizás el primer uso de grupos control en psicología y educación ocurrió en 1901 con la publicación de Thorndike y Woodworth (1901). Uno de los hombres que realizó esta investigación, Thorndike, extendió las ideas básicas y revolucionarias de esta primera investigación a la educación (Thorndike, 1924). En el gigantesco estudio de Thorndike de 8 564 alumnos de muchas escuelas en varias ciudades, los controles fueron grupos educativos independientes. Entre otras comparaciones, él contrastó las ganancias en las puntuaciones de pruebas de inteligencia, presumiblemente generadas por el estudio de inglés, historia, geometría y latín con las presumiblemente generadas por el estudio de inglés, historia, geometría y taller. En efecto, él intentó comparar la influencia del latín y del

taller. También realizó otras comparaciones de naturaleza similar. A pesar de la debilidad del diseño y del control, sus experimentos y otros realizados por quienes él mismo estimuló, eran excepcionales por su discernimiento. Thorndike incluso criticó a sus colegas por no admitir estudiantes de estenografía y mecanografía que no habían estudiado latín, debido a que él reclamaba haber demostrado que la influencia de otros factores sobre la inteligencia era similar. Es interesante el hecho de que él pensara que se necesitaba un número enorme de participantes —exigió 18 000 casos más—. En 1924 estaba bastante consciente de la necesidad de emplear muestras aleatorias.

La noción del grupo control requiere generalización. Suponga que en un experimento educativo se tienen cuatro grupos experimentales como siguen. En A_1 se da un reforzamiento por cada respuesta, en A_2 se da en intervalos regulares de tiempo, en A_3 se da en intervalos aleatorios, y en A_4 no se da el reforzamiento. Técnicamente hay tres grupos experimentales y un grupo control, en el sentido tradicional del grupo control. Sin embargo, A_4 podría ser otro "tratamiento experimental"; podría ser algún tipo de reforzamiento mínimo. Entonces, en el sentido tradicional no habría grupo control. El sentido tradicional del término "grupo control" carece de generalidad. Si se generaliza el concepto de grupo control, la dificultad desaparece. Siempre que haya más de un grupo experimental y a cualesquiera dos grupos se les apliquen diferentes tratamientos, el control está presente en el sentido comparativo antes mencionado. Mientras exista un intento por hacer a dos grupos sistemáticamente diferentes en una variable dependiente, será posible una comparación. Por lo tanto, el concepto tradicional de que un grupo experimental debe recibir el tratamiento, y que éste no se da a un grupo control, es un caso especial de la regla más general de que se necesitan grupos de comparación para la validez interna de la investigación científica.

Si el razonamiento es correcto, es posible establecer diseños como el siguiente:

	X_1	Y
$[A]$	X_2	Y
	X_3	Y

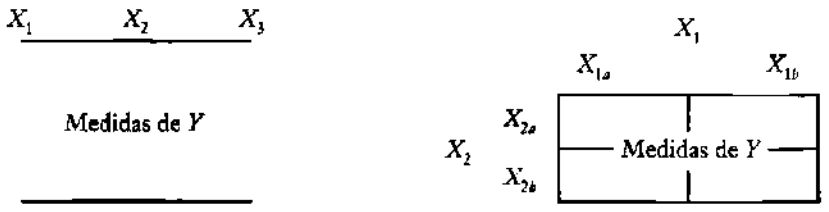
o

	X_{1a}	Y
$[A]$	X_{1b}	Y
	X_{2a}	Y
	X_{2b}	Y

Estos diseños se reconocerán más fácilmente si se construyen a la manera del análisis de varianza, como en la figura 20.3. El diseño de la izquierda es un diseño simple de análisis de varianza de un factor; y el de la derecha, un diseño factorial de $2 \times 3 \times 2$. En el diseño del lado derecho, X_{1a} puede ser el grupo experimental y X_{1b} el control, con X_{2a} y X_{2b} como una variable manipulada o una variable atributiva dicotómica. Éste es, por supuesto, el mismo diseño que se presenta en la figura 20.2(*a*).

La estructura del diseño 20.2 es la misma que aquella del diseño 20.1. La única diferencia es que los participantes están apareados en uno o más atributos. Sin embargo, para

FIGURA 20.3



que el diseño sea un diseño "adecuado", la aleatorización debe aparecer en escena, como se nota por la *a* unida a la *Ap* (de "apareado"). No es suficiente utilizar participantes apareados. Los miembros de cada par deben asignarse aleatoriamente a los dos grupos. Idealmente, el que un grupo sea el grupo experimental o el control debe decidirse también en forma aleatoria. En cualquier caso, cada decisión ha de tomarse lanzando una moneda o utilizando una tabla de números aleatorios; los números pares se usan para un grupo y los nones para el otro. Si existen más de dos grupos, debe utilizarse un sistema de números aleatorios.

Diseño 20.2: Grupo experimental-grupo: participantes apareados

$[Ap_a]$	X	Y	(Experimental)
	$\sim X$	Y	(Control)

Como en el diseño 20.1, es posible, aunque no siempre fácil, utilizar más de dos grupos. (La dificultad de aparear más de dos grupos se estudió anteriormente.) Sin embargo, hay ocasiones en que un diseño apareado constituye un elemento inherente a la situación de investigación. Cuando se utilizan los mismos participantes para dos o más tratamientos experimentales, o cuando se les aplica más de un ensayo a los participantes, el apareamiento es inherente a la situación. En investigación educativa, las escuelas o clases son efectivamente variables, cuando, digamos, se utilizan dos o más escuelas o clases y los tratamientos experimentales se administran en cada escuela o clase, entonces el diseño 20.2 constituye la base de la lógica del diseño. Estudie el paradigma de un diseño de escuelas en la figura 20.4. Hay que destacar el hecho de que la varianza debida a las diferencias entre escuelas —y es posible que dicha varianza sea sustancial— puede ser estimada de inmediato.

Apareamiento contra aleatorización

Aunque la aleatorización, que incluye la selección aleatoria y la asignación aleatoria, constituye el método preferido para controlar la varianza extraña, el uso del apareamiento también tiene su mérito. En varias situaciones fuera de los círculos académicos, los científicos del comportamiento no serán capaces de utilizar la aleatorización para lograr constancia entre los grupos, antes de la administración del tratamiento. En una universidad por lo común se dispone de un universo de participantes de donde seleccionar a los participantes. Los investigadores en una situación como ésta pueden darse el gusto de utilizar procedimientos aleatorios. Sin embargo, en la investigación de negocios éste quizá no sea el caso.

FIGURA 20.4

Escuelas	X_{e1} Experimental 1	X_{e2} Experimental 2	X_c Control
1			
2			
3		Medidas de Y	
4			
5			

Entre los investigadores de mercado es popular la prueba de tienda controlada, que consiste en un experimento de campo. El segundo autor ha conducido dichos estudios para varias compañías de investigación de mercado y para una cadena de tiendas de abarrotes en el sur de California. Una de las metas de la prueba de tienda controlada es ser muy discreto. Si un fabricante de productos de jabón quiere determinar los efectos de un cupón de descuento en la conducta de compra del consumidor, el fabricante no desea que el fabricante competidor de un producto similar se entere del asunto. ¿Por qué? Porque si un competidor supiera que se está realizando un estudio de investigación en una tienda, podría ir y acaparar su propio producto, contaminando así el estudio.

Para regresar al análisis sobre la aleatorización *versus* el apareamiento, muchas veces una cadena de tiendas de abarrotes o una cadena de tiendas departamentales tiene un número finito de tiendas para utilizar en un estudio. La ubicación y la clientela ejercen una enorme influencia sobre las ventas. Generalmente las ventas son la variable dependiente de dichos estudios. Con un número limitado de tiendas de donde escoger para realizar la investigación, la asignación aleatoria a menudo no funciona para igualar grupos de tiendas. Una tienda de la cadena puede hacer tres o cuatro veces el volumen de transacciones que otra y si es elegida al azar creará un gran desequilibrio en el grupo al que pertenezca, especialmente si el otro grupo no incluye una tienda similar para equilibrarlo; en pocas palabras, los grupos ya no serán iguales. Por lo tanto, la solución aquí consiste en aparear las tiendas con una base individual; un miembro del par se asigna aleatoriamente a una condición experimental y el otro miembro recibe la otra condición. Con más de dos condiciones, más tiendas tendrían que ser apareadas y luego asignarse a las condiciones de tratamiento.

En algunos estudios de ingeniería de factores humanos usando simuladores, el uso de la aleatorización en ocasiones no es factible económica ni prácticamente. Considere la prueba de dos configuraciones para un simulador. Un investigador desea saber cuál conduce a menos errores perceptuales. Los procesos de aleatorización dirían que el investigador debería asignar a los participantes aleatoriamente a las condiciones, conforme entren en el estudio. Sin embargo, cuando se requieren de tres a seis meses para cambiar la configuración del simulador, ya no es factible proceder de la manera "usual".

Un punto importante a recordar es que la aleatorización —cuando puede llevarse a cabo correcta y apropiadamente— por lo común es mejor que el apareamiento. Es quizás el único método para controlar fuentes de varianza desconocidas. Uno de los principales inconvenientes del apareamiento es que no se puede estar seguro de que se ha realizado un par exacto. Sin esa precisión, la inexactitud puede ser una explicación alternativa de por qué la variable dependiente es diferente entre las condiciones de tratamiento, después de la manipulación experimental.

que se desea tener dos o más grupos apareados por inteligencia, y que se quiere utilizar el método de apareamiento de distribución de frecuencias. Primero se necesitará una puntuación de una prueba de inteligencia para cada niño. Después se requiere crear los dos o más grupos de tal manera que los grupos tengan la misma puntuación promedio en la prueba de inteligencia, así como la misma desviación estándar y una simetría o asimetría semejante en las puntuaciones. Cada grupo sería estadísticamente igual —la media, la desviación estándar y la simetría o asimetría entre cada grupo sería estadísticamente equivalente—. Podría utilizarse una prueba estadística de hipótesis; pero el investigador requiere estar consciente de que es necesario considerar los dos tipos de error. Si más de una variable se considerara relevante para aparear a los participantes, entonces se requeriría que cada grupo de participantes tuviera las mismas medidas estadísticas en todas estas variables. El número de participantes perdidos al utilizar dicha técnica no sería tan grande como el número de pérdidas al utilizar el método de individuo por individuo, ya que cada participante adicional tan sólo tendría que contribuir para producir las medidas estadísticas apropiadas, en lugar de ser idéntico a otro participante en las variables relevantes. Por lo tanto, esta técnica es más flexible en términos de capacidad para utilizar a un participante particular.

La principal desventaja de realizar el apareamiento mediante el método de distribución de frecuencias, sucede cuando se aparea con base en más de una variable. Aquí la combinación de variables puede estar mal apareada en los diversos grupos. Si se fueran a aparear edad y tiempo de reacción, un grupo podría incluir participantes mayores con tiempos de reacción más lentos, y participantes más jóvenes con tiempos de reacción más rápidos, mientras que el otro grupo podría tener la combinación opuesta. La media y distribución de las dos variables sería equivalente; pero los participantes en cada grupo serían completamente diferentes. Tal diferencia podría afectar a la variable dependiente.

Apareamiento mediante mantener constantes las variables

Otra técnica que se utiliza para crear grupos igualados de participantes consiste en mantener constante la variable extraña o no planeada. Todos los participantes en cada grupo experimental tendrán el mismo grado o tipo de variable extraña. Si se necesita controlar la variación causada por diferencias de género, se puede mantener constante el género utilizando únicamente hombres o únicamente mujeres en el estudio. Esto tiene el efecto de aparear a todos los participantes en términos de la variable género. Este procedimiento de apareamiento crea una muestra de participantes más homogénea, debido a que solamente se utilizan participantes con cierto tipo o cantidad de la variable fortuita. Muchos proyectos de investigación de estudiantes en universidades utilizan este método, especialmente cuando el universo de participantes posee una mayoría de hombres o de mujeres. Esta técnica de mantener constantes las variables tiene por lo menos dos problemas que podrían afectar la validez del estudio. La severidad del problema se incrementa si se mantienen constantes demasiadas variables. La primera desventaja consiste en que la técnica restringe el tamaño de la población de participantes. Como consecuencia, en algunos casos resulta complicado encontrar suficientes participantes para incluir en el estudio. La investigación pionera de seccionar el cerebro, realizada por Roger Sperry, con frecuencia ha sido criticada por la restricción de los participantes utilizados en el estudio. Sus primeros trabajos incluyeron solamente pacientes epilépticos. Así, un estudio que utilice dicho método podría ser criticado por tener un sesgo en la selección.

La segunda desventaja es aún más crítica, ya que los resultados sólo pueden generalizarse al tipo de participante utilizado en el estudio. Los resultados obtenidos del estudio

de los pacientes epilépticos sólo podían generalizarse a otros pacientes epilépticos. Si alguien deseara saber si pacientes no epilépticos experimentarían los mismos cambios perceptuales, el investigador tendría que conducir un estudio similar con pacientes no epilépticos. Las conclusiones de dicho estudio podrían, en realidad, ser iguales a los obtenidos en el estudio con pacientes epilépticos, pero deben realizarse dos estudios separados. La única manera de averiguar si los resultados de un estudio pueden generalizarse a la población es replicando el estudio con participantes de diferentes características.

Apareamiento mediante la incorporación de una variable extraña al diseño de investigación

Otra forma para intentar desarrollar grupos igualados es utilizar la variable extraña como variable independiente en el diseño de investigación. Suponga que se conduce un experimento de aprendizaje con ratas y que se desean controlar los efectos del peso. La idea aquí es que el animal con mayor peso necesitará ingerir más alimento después de un periodo de privación y, por lo tanto, estará más motivada. Si se hubiese utilizado el método de mantener constante el peso, se tendrían muchos menos participantes. Al utilizar el peso como variable independiente se pueden utilizar muchos más participantes en el estudio. En términos estadísticos, un aumento en el número de participantes significa un aumento en poder y sensibilidad. Con el uso de una variable extraña como variable independiente en el diseño, se aísla una fuente de varianza sistemática y también se determina si la variable extraña tiene un efecto sobre la variable dependiente.

Sin embargo, la incorporación de una variable extraña en el diseño no debe efectuarse indiscriminadamente. Lograr que la variable extraña forme parte del diseño de investigación parece ser un excelente método de control; pero dicho método está mejor utilizado cuando existe un interés en las diferencias producidas por la variable extraña, o en la interacción entre la variable extraña y otras variables independientes. El investigador puede incluso incorporar al diseño una variable medida en una escala continua. La diferencia entre una variable extraña continua y una discreta residiría en la etapa del análisis de datos del proceso de investigación. Entonces sería preferible el uso de la regresión múltiple o del análisis de covarianza, en lugar del análisis de varianza.

Los participantes como su propio control

Puesto que cada individuo es único, resulta difícil, si no imposible, encontrar a otro individuo que fuera el par perfecto. Sin embargo, una sola persona es siempre un perfecto par para sí misma. Una de las técnicas más poderosas para lograr la igualdad y la constancia de los grupos experimentales, antes de la administración del tratamiento, consiste en utilizar a esa misma persona en cada condición del experimento. Algunos se refieren a lo anterior como el uso de participantes como su propio control. Aparte de la reactividad del experimento en sí, la posibilidad de que surja variación extraña debida a diferencias entre individuos se minimiza drásticamente. Tal método para lograr la constancia es común en algunas áreas de las ciencias del comportamiento. En psicología, el estudio de la interfase de seres humanos y máquinas (factores humanos o ingeniería humana) utiliza este método. Simon (1976) presenta varios diseños experimentales interesantes que utilizan al mismo participante en muchas condiciones de tratamiento. Sin embargo, dicho método no se ajusta a todas las aplicaciones. Algunos estudios relacionados con el aprendizaje no son elegibles, pues una persona no puede *desaprender* un problema para poderle aplicar un

método diferente después. El uso de este método requiere de mayor planeación y de una ejecución más precisa que otros métodos.

Extensiones adicionales del diseño: diseño 20.3 utilizando un pretest

El diseño 20.3 tiene muchas ventajas y se utiliza con frecuencia. Su estructura es similar a la del diseño 19.2, con dos diferencias importantes: el diseño 19.2 carece de grupo control y de aleatorización. El diseño 20.3 es similar a los diseños 20.1 y 20.2, excepto en que se añadió la situación "antes" o de pretest. Con frecuencia se utiliza para estudiar cambios. Como los diseños 20.1 y 20.2, el diseño 20.3 puede expandirse a más de dos grupos.

Diseño 20.3: grupo control antes y después (pretest-postest)

(*a*)	[*A*]	Y_b	X	Y_a	(Experimental)
		Y_b	$\sim X$	Y_a	(Control)
(*b*)	[Ap_a]	Y_b	X	Y_a	(Experimental)
		Y_b	$\sim X$	Y_a	(Control)

En el diseño 20.3(*a*) los participantes son asignados aleatoriamente al grupo experimental (línea superior) y al grupo control (línea inferior), y son sometidos a una situación de pretest en una medida de Y, la variable dependiente. Entonces, el investigador puede verificar la igualdad de los dos grupos respecto a Y. La manipulación experimental X se lleva a cabo y, después, los grupos son medidos nuevamente respecto a Y. La diferencia entre los dos grupos se prueba estadísticamente. Una característica interesante y difícil de este diseño es la naturaleza de las puntuaciones que se analizan generalmente: puntuaciones de diferencia o de cambio, $Y_d - Y_a = D$. A menos que el efecto de la manipulación experimental sea fuerte, no se recomienda el análisis de las puntuaciones de diferencia. Las puntuaciones de diferencia son considerablemente menos confiables que las puntuaciones a partir de las cuales se calculan. Una explicación clara del porqué de lo anterior la ofrecen Friedenberg (1995) y Sax (1997). Aunque existen otros problemas, aquí se analizan sólo las principales fortalezas y debilidades (para un estudio más completo al respecto, véase Campbell y Stanley, 1963). Al final del análisis se examinarán las dificultades analíticas de las puntuaciones de diferencia o de cambio.

Quizás de mayor importancia, el diseño 20.3 supera la gran debilidad del diseño 19.2, ya que brinda un grupo control contra el cual puede verificarse la diferencia, $Y_d - Y_a$. Con sólo un grupo no es posible saber si la historia, la maduración (o ambas), o la manipulación experimental X produjeron el cambio en Y. Cuando se añade un grupo control, la situación se altera radicalmente. Después de todo, si se igualan los grupos (a través de la aleatorización), los efectos de la historia y de la maduración, en caso de estar presentes, deberían presentarse en ambos grupos. Si se incrementan las edades mentales de los niños del grupo experimental, también deberían hacerlo así las edades mentales de los niños del grupo control. Entonces, si aún existe una diferencia entre las medidas de Y de los dos grupos, ésta no debe ser a causa de la historia o de la maduración; es decir, si algo afecta a los participantes del grupo experimental entre el pretest y el postest, ese algo también debería afectar a los participantes del grupo control. De manera similar, el efecto de realizar una prueba —medidas reactivas de Campbell— debe controlarse porque, si la prueba afecta a los miembros del grupo experimental, también debe afectar de forma similar a los

miembros del grupo control. (Sin embargo, existe una debilidad encubierta aquí, que se estudiará más adelante.) Ésta es la principal fuerza del diseño de grupo control-grupo experimental con pretest-postest, bien planeado y bien ejecutado.

Por otro lado, los diseños de pretest-postest tienen un aspecto problemático que disminuye tanto la validez interna como la validez externa del experimento. Esta fuente de dificultad es el pretest. Un pretest puede tener un efecto sensibilizador en los participantes. Respecto a la validez interna, por ejemplo, los participantes serían alertados sobre ciertos eventos en su ambiente, que podrían no haber notado comúnmente. Si el pretest consiste en una escala de actitud, tal vez sensibilice a los participantes respecto a los aspectos o problemas mencionados en la escala; así, cuando se administre el tratamiento X al grupo experimental, los participantes de este grupo pueden responder no tanto a la influencia tentativa (la comunicación o cualquier método que se utilice para cambiar actitudes), sino a la combinación de su sensibilidad incrementada respecto al tema y a la manipulación experimental.

Puesto que dichos efectos de interacción no son inmediatamente obvios, y como representan una amenaza para la validez externa de los experimentos, vale la pena considerarlos un poco más. Se pensaría que, puesto que tanto al grupo control como al grupo experimental se les aplica el pretest, el efecto de prueba previa, si acaso hay alguno, aseguraría la validez del experimento. Suponga que no se realiza ningún pretest, es decir, que se utiliza el diseño 20.2. Si lo demás permanece igual, una diferencia entre los grupos experimental y control después de la manipulación experimental de X puede asumirse como efecto de X. No existe razón alguna para suponer que un grupo es más sensible o que está más alerta que el otro, ya que ambos enfrentaron la situación de prueba después de X. Pero cuando se utiliza un pretest, la situación cambia. Mientras que el pretest sensibiliza a ambos grupos, puede hacer que los participantes experimentales respondan a X, completa o parcialmente, debido a la sensibilidad.

También se tiene una carencia de generalización o validez externa, ya que puede ser posible generalizar a los grupos probados antes; pero no a los grupos que no fueron probados antes. En efecto, esta situación molesta al investigador, porque: ¿quién quiere generalizar a los grupos probados antes?

Si esta debilidad se vuelve importante, ¿por qué éste es un buen diseño? Mientras que el posible efecto de interacción descrito antes repercutiría en alguna investigación, es dudoso que afecte fuertemente a la mayoría de la investigación del comportamiento, si los investigadores están conscientes de su potencial y toman precauciones adecuadas. Realizar pruebas constituye una parte normal y aceptada en muchas situaciones, especialmente en educación. Por lo tanto, resulta dudoso que los participantes de la investigación estén excesivamente sensibilizados en dichas situaciones. Aun así, pueden existir ocasiones en que resulten afectados. La regla dada por Campbell y Stanley (1963) es buena: cuando se van a utilizar procedimientos de prueba inusuales, es mejor usar diseños sin pretest.

Puntuaciones de diferencia

Observe el diseño 20.3 otra vez, particularmente los cambios entre Y_a y Y_d. Uno de los problemas más difíciles que ha abrumado e intrigado a los investigadores, a los especialistas en medición y a los estadísticos es cómo estudiar y analizar dichas puntuaciones de diferencia o de cambio. En un libro con el alcance de éste, sería imposible entrar a los problemas con detalle. El lector interesado puede leer dos excelentes libros editados: el de Harris (1963) y el de Collins y Horn (1991). No obstante, se esbozarán algunos preceptos y precauciones generales. Podría pensarse que la aplicación del análisis de varianza a las puntuaciones de

diferencia producidas por el diseño 20.3 y diseños similares, sería efectiva. Dicho análisis puede realizarse si los efectos experimentales son sustanciales. Pero las puntuaciones de diferencia, como se mencionó antes, son por lo general menos confiables que las puntuaciones a partir de las cuales se calculan. Las diferencias reales entre los grupos experimental y control quizá no sean detectables simplemente por la baja confiabilidad de las puntuaciones de diferencia. Para detectar las diferencias entre los grupos control y experimental, las puntuaciones analizadas deben ser lo suficientemente confiables para reflejar las diferencias y, así, ser detectables por medio de pruebas estadísticas. Debido a tal dificultad, investigadores como Cronbach y Furby (1970) dicen que las puntuaciones de diferencia o cambio no deben utilizarse. ¿Entonces, qué se puede hacer?

El procedimiento recomendado por lo común consiste en usar las llamadas puntuaciones residualizadas o regresionadas de ganancia, las cuales se calculan al predecir las puntuaciones del postest a partir de las puntuaciones del pretest, con base en la correlación entre el pretest y el postest, y sustrayendo después estas puntuaciones predichas de las puntuaciones del postest, para obtener las puntuaciones residuales de ganancia. (El lector no debe preocuparse si este procedimiento no está demasiado claro en este momento. Más tarde, después del estudio de la regresión y del análisis de covarianza deberá quedar más claro.) El efecto de las puntuaciones del pretest se elimina de las puntuaciones del postest, es decir, las puntuaciones residuales son puntuaciones del postest purificadas respecto a la influencia del pretest. Después se prueba la significancia de la diferencia entre las medias de estas puntuaciones. Todo lo cual puede lograrse utilizando el procedimiento descrito y una ecuación de regresión, o por medio de un análisis de covarianza.

Sin embargo, incluso el uso de puntuaciones residuales de ganancia y el análisis de covarianza no son perfectos. Si los participantes no han sido asignados aleatoriamente a los grupos experimental y control, el procedimiento no salvará la situación. Cronbach y Furby (1970) señalan que cuando los grupos difieren sistemáticamente antes del tratamiento experimental, en otras características pertinentes a la variable dependiente, la manipulación estadística no corrige tales diferencias. Sin embargo, si se utiliza un pretest, es mejor usar la asignación aleatoria y el análisis de covarianza, recordando que los resultados siempre deben tratarse con especial cuidado. Por último, el análisis de regresión múltiple proporciona la mejor solución para el problema, como se verá más adelante. Por desgracia las complejidades del diseño y del análisis estadístico pueden desanimar al estudiante de investigación, incluso al punto de hacerlo sentir desamparado. No obstante, así es la naturaleza de la investigación del comportamiento: tan sólo refleja el carácter excesivamente complejo de la realidad psicológica, sociológica y educativa, lo cual resulta frustrante y emocionante al mismo tiempo; como el matrimonio, la investigación del comportamiento es difícil y con frecuencia poco exitosa, pero no imposible. Además, es una de las mejores formas de adquirir un entendimiento confiable de nuestro mundo del comportamiento. El punto de vista de este libro es que se debe aprender y comprender lo más posible sobre lo que hacemos, que se debe tener un cuidado razonable con el diseño y el análisis, y que hay que realizar la investigación sin preocuparse demasiado sobre los aspectos analíticos. La cuestión principal es siempre el problema de investigación y el interés que se tenga en él. Ello no implica una desatención o desprecio al análisis; significa simplemente un entendimiento y cuidado razonables, y cantidades sanas tanto de optimismo como de escepticismo.

Diseño 20.4: Simulación antes-después, aleatorizado

[*A*]		X	Y_a
	Y_b		

El valor del diseño 20.4 es dudoso, aunque se considera un diseño adecuado. La demanda científica de realizar una comparación está satisfecha; hay un grupo de comparación (línea inferior). Una debilidad importante del diseño 19.3 (una versión débil del diseño 20.4) se soluciona por medio de la aleatorización. Recuerde que con el diseño 19.3 era imposible suponer de antemano que los grupos experimental y control eran equivalentes. El diseño 20.4 requiere que los participantes sean asignados aleatoriamente a ambos grupos; entonces es posible suponer que estadísticamente son iguales. Un diseño de este tipo puede utilizarse cuando existe la preocupación del efecto reactivo del pretest; o cuando, a causa de las exigencias de situaciones prácticas, no se tiene otra opción. Dicha situación sucede cuando se tiene una única oportunidad de probar un método o alguna innovación. Para probar la eficacia del método, se proporciona una línea base para juzgar el efecto de X sobre Y, aplicando un pretest a un grupo similar al grupo experimental. Entonces Y_d se prueba contra Y_a.

La validez de este diseño se desploma si los dos grupos no son seleccionados aleatoriamente a partir de la misma población, o si los participantes no son asignados aleatoriamente a los dos grupos. Además, incluso si se utiliza la aleatorización, no existe una garantía real de que ésta funcionó para igualar los dos grupos antes del tratamiento. Comparte con otros diseños similares las debilidades mencionadas, es decir, otras posibles variables pueden influir en el intervalo comprendido entre Y_a y Y_d. En otras palabras, el diseño 20.4 es superior al diseño 19.3; pero no debe utilizarse si está disponible otro diseño que se considere mejor.

Diseño 20.5: Tres grupos, antes-después

	Y_a	X	Y_d	(Experimental)
[A]	Y_a	$\sim X$	Y_d	(Control 1)
		X	Y_d	(Control 2)

El diseño 20.5 es mejor que el diseño 20.4. Además de las ventajas del diseño 20.3, proporciona una forma posible de evitar la confusión debida a los efectos interactivos del pretest. Esto se logra por medio de un segundo grupo control (tercera línea). (Parece un poco extraño tener un grupo control con X; pero el grupo de la tercera línea es realmente un grupo control.) El hecho de contar con las medidas de Y_d de este grupo, hace posible verificar el efecto de interacción. Suponga que la media del grupo experimental es significativamente mayor que la media del grupo control 1. Podría dudarse si tal diferencia se debe realmente a X; quizá se produjo por un incremento en la sensibilidad de los participantes después del pretest y por la interacción de su sensibilidad y X. Ahora observe la media de Y_d del grupo control 2. Ésta también debería ser significativamente mayor que la media del grupo control 1. Si es así, se puede suponer que el pretest no sensibilizó excesivamente a los participantes, o que X es lo suficientemente fuerte para anular un efecto de interacción entre la sensibilización y X.

Diseño 20.6: Cuatro grupos, antes-después (Solomon)

	Y_a	X	Y_d	(Experimental)
[A]	Y_a	$\sim X$	Y_d	(Control 1)
		X	Y_d	(Control 2)
		$\sim X$	Y_d	(Control 3)

El diseño propuesto por Solomon (1949) es fuerte y estéticamente satisfactorio. Posee controles potentes. En realidad, si se cambia la designación de control 2 por experimental 2, se tiene una combinación de los diseños 20.3 y 20.1, los dos mejores diseños, donde el primer diseño forma las primeras dos líneas del diseño de Solomon; y el último, las segundas dos líneas. Las virtudes de ambos se combinan en un diseño. Aunque este diseño puede tener una forma apareada, eso no se analiza aquí ni se recomienda su uso. Campbell (1957) afirma que tal diseño se ha convertido en un nuevo ideal para los científicos sociales. Aunque parece una afirmación demasiado fuerte, indica la alta estima en que se tiene al diseño.

Una de las razones de la fuerza del diseño es que la demanda de comparación queda bien satisfecha con las dos primeras líneas y con las segundas dos líneas. La aleatorización incrementa la probabilidad de la equivalencia estadística de los grupos; y la historia y la maduración están controladas por las primeras dos líneas del diseño. El efecto de interacción debido a la posible sensibilización por el pretest en los participantes está controlado por las primeras tres líneas. Al añadir la cuarta línea, se controlan los efectos temporales contemporáneos que pueden haber ocurrido entre Y_1 y Y_2. Ya que los diseños 20.2 y 20.3 están combinados, se tiene la fuerza de cada uno por separado y la fuerza de la replicación porque en efecto, hay dos experimentos. Si Y_2 del grupo experimental es significativamente mayor que la del grupo control 1, y la del grupo control 2 es significativamente mayor que la del grupo control 3, aunado a la consistencia de los resultados de ambos experimentos, entonces ésta es una fuerte evidencia de la validez de la hipótesis de investigación.

¿Qué defecto puede tener este ideal de diseño? Ciertamente luce bien en el papel. Parece haber solamente dos fuentes de debilidad. Una es de tipo práctico ya que es más difícil realizar dos experimentos simultáneamente, que uno; y el investigador se encuentra con la dificultad de localizar más participantes del mismo tipo.

La otra dificultad es estadística. Observe que hay una falta de balance entre los grupos. Existen cuatro grupos, pero no hay cuatro conjuntos completos de medidas. Utilizando las primeras dos líneas, es decir, con el diseño 20.3, se puede sustraer Y_1 de Y_2 o hacer un análisis de covarianza. Con las dos líneas se pueden probar las Y_2 contra sí mismas con una prueba t o una prueba F; pero el problema reside en cómo obtener un enfoque estadístico general. Una solución es probar las Y_2 de los grupos control 2 y 3 contra el promedio de las dos Y_2 (las primeras dos líneas), así como también probar la significancia de la diferencia de las Y_2 de las primeras dos líneas. Además, Solomon originalmente sugirió un análisis factorial de varianza de 2×2, utilizando los cuatro conjuntos de medidas de Y_2. La sugerencia de Solomon se presenta en la figura 20.5. Un estudio cuidadoso revelará que éste es un buen ejemplo de pensamiento de investigación, una excelente combinación de diseño y análisis. Con este análisis se pueden estudiar los efectos principales, X y $\sim X$, y aquellos con pretest y sin pretest. Lo que es más interesante, se puede probar la interacción de la prueba previa y X, y obtener una respuesta clara al problema anterior.

Mientras que éste y otros diseños complejos tienen fuerzas notorias, es dudoso que puedan utilizarse rutinariamente. De hecho, quizá deban reservarse para experimentos muy importantes, en los cuales, se prueben de nuevo con mayor rigor y control, hipótesis ya probadas con diseños más simples. En efecto, se recomienda que diseños como el 20.5 y el 20.6 y ciertas variantes del diseño 20.6 (que se discutirá más adelante) se reserven para

FIGURA 20.5

	X	$\sim X$
Con pretest	X_2, experimental 1	Y_2, control 1
Sin pretest	Y_2, control 2	Y_2, control 3

pruebas definitivas de hipótesis de investigación, después de que se haya realizado cierta cantidad de experimentación previa.

RESUMEN DEL CAPÍTULO

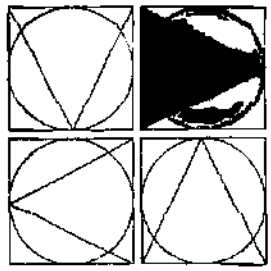
1. El diseño de un estudio es el proyecto o plan para desarrollar la investigación.
2. Un diseño es un subconjunto de un producto cartesiano cruzado de varios niveles de la variable independiente.
3. El diseño experimental es aquel donde se manipula al menos una de las variables independientes utilizadas en el estudio.
4. Los diseños no experimentales son aquellos en los cuales no hay aleatorización para igualar los grupos antes de administrar los tratamientos.
5. Generalmente el método estadístico más apropiado para los diseños experimentales es el análisis de varianza.
6. Los supuestos del análisis de varianza con frecuencia se violan en los diseños no experimentales. La regresión múltiple puede ser un método de análisis de datos más apropiado para diseños no experimentales.
7. El diseño de grupo experimental-grupo control con participantes aleatorizados (diseño 20.1) es el mejor para muchos estudios de investigación experimentales del comportamiento.
8. El diseño de cuatro grupos de Solomon (diseño 20.6) maneja varias de las preocupaciones de la investigación del comportamiento. Sin embargo, utiliza los recursos de dos estudios y quizá no resulte eficiente económicamente.
9. El diseño 20.2 es como el diseño 20.1, excepto que utiliza participantes apareados.
10. El uso de participantes apareados se vuelve útil en situaciones donde la aleatorización no funciona apropiadamente.
11. Existen varias formas de aparear participantes. La más popular es el método de individuo por individuo.
12. El apareamiento tiene problemas en cuanto a que el investigador nunca puede estar seguro de que todas las variables importantes hayan sido utilizadas en el proceso. Además, si se utilizan demasiadas variables en el apareamiento, se vuelve más difícil encontrar participantes con características iguales.
13. El diseño 20.3 utiliza un pretest; cuya aplicación es una forma para determinar si los grupos son iguales o si funcionó la aleatorización. Sin embargo, aplicar un pretest también sensibiliza a los participantes del experimento.
14. Las puntuaciones de diferencia se utilizan con frecuencia en diseños que incluyen un pretest. Sin embargo, las puntuaciones de diferencia pueden no ser confiables.
15. El diseño 20.4 es un diseño de antes-después simulado que utiliza participantes aleatorizados. El segundo grupo (control) tan sólo se mide en un pretest. El grupo experimental recibe el tratamiento y el postest.
16. El diseño 20.5 es un diseño de tres grupos antes-después. Es como el diseño 20.3, excepto por la introducción de un tercer grupo que recibe tratamiento y porque no se utiliza un pretest.

SUGERENCIAS DE ESTUDIO

1. La primera oración de este capítulo fue: "El diseño constituye una disciplina de datos." ¿Qué significa dicha frase? Justifique su respuesta.

2. Suponga que usted es un psicólogo educativo y planea probar la hipótesis de que retroalimentar con información psicológica a los maestros, incrementa el aprendizaje de los niños, al aumentar el entendimiento del maestro sobre los niños. Bosqueje un diseño de investigación ideal para probar esta hipótesis, suponiendo que usted tiene un completo dominio de la situación y suficiente dinero y asistencia. (Éstas son condiciones importantes, que se incluyen para liberar al lector de las complicaciones prácticas que tan frecuentemente comprometen los buenos diseños de investigación.) Establezca dos diseños, cada uno con aleatorización completa; que ambos diseños sigan el paradigma del diseño 20.1. En uno de ellos utilice sólo una variable independiente y un análisis de varianza de un factor. En el segundo, emplee dos variables independientes y un diseño factorial simple. ¿En qué difieren estos dos diseños respecto a sus poderes de control y en la información que producen? ¿Cuál prueba mejor la hipótesis? Explique por qué.
3. La recomendación del texto de no utilizar el análisis de varianza en investigación no experimental no aplica en gran medida al análisis de varianza de un factor ni al análisis factorial. Tampoco aplica el problema del mismo número de casos en las casillas. De hecho, en varios estudios no experimentales, el análisis de varianza de un factor se ha utilizado provechosamente. Uno de dichos estudios es el de Jones y Cook (1975). La variable independiente era la actitud hacia los afroamericanos, obviamente no manipulada. La variable dependiente era el grado de apoyo hacia una política social que afectaba a los afroamericanos.

   Se sugiere que los lectores revisen y asimilen este estudio. Usted quizás también desee realizar un análisis de varianza con los datos de la tabla 1 de los autores, utilizando el método del análisis de varianza con el uso de n, medias y desviaciones estándar, descrito anteriormente (véase anexo, capítulo 13).



CAPÍTULO 21

APLICACIONES DEL DISEÑO DE INVESTIGACIÓN: GRUPOS ALEATORIZADOS Y GRUPOS CORRELACIONADOS

Resulta difícil explicarle a alguien cómo hacer investigación. Quizás lo mejor sea asegurarse de que el principiante capte los principios y las posibilidades. Además los enfoques y las

tácticas pueden sugerirse. Para abordar un problema de investigación, el investigador debe dejar volar la mente, especular acerca de las posibilidades e incluso adivinar el patrón de los resultados. Una vez que se conocen las posibilidades, las intuiciones pueden seguirse y explorarse. Sin embargo, la intuición y la imaginación no son de mucha ayuda si se sabe poco o nada acerca de recursos técnicos. Por otro lado, la buena investigación no consiste sólo de metodología y técnica; el pensamiento intuitivo es esencial porque ayuda a los investigadores a alcanzar soluciones que no son convencionales o rutinarias. No obstante, nunca debe olvidarse que el pensamiento analítico y el pensamiento intuitivo creativo dependen del conocimiento, el entendimiento y la experiencia.

Los principales propósitos de este capítulo consisten en enriquecer e ilustrar el diseño y la discusión estadística con ejemplos reales de investigación, así como sugerir posibilidades básicas para diseñar investigación para que el estudiante finalmente resuelva problemas de investigación. El propósito general es, entonces, complementar y enriquecer análisis estadísticos y de diseño previos y más abstractos.

Diseño simple de sujetos aleatorizados

En los capítulos 13 y 14 se estudiaron e ilustraron los estadísticos del análisis de varianza simple de un factor y del análisis factorial de varianza. El diseño detrás de la discusión anterior se llama *diseño de sujetos aleatorizados*. El paradigma general del diseño (denominado como diseño 20.1) se presenta a continuación:

[*A*]	X	Y	(Experimental)
	$\sim X$	Y	(Control)

Ejemplo de investigación

La forma más simple del diseño 20.1 es un paradigma del análisis de varianza de un factor, en el cual a k grupos se les dan k tratamientos experimentales, y las k medias se comparan mediante un análisis de varianza o con pruebas separadas de significancia. Un vistazo a la parte izquierda de la figura 20.3 presenta esta forma simple del diseño 20.1 con $k = 3$. Aunque parezca extraño, no se utiliza muy a menudo; los investigadores prefieren la forma factorial del diseño 20.1 con mayor frecuencia. Más adelante se presenta un ejemplo de un factor; se utiliza la asignación aleatoria. Por desgracia algunos investigadores no reportan cómo se asignaron los participantes a los grupos o a los tratamientos. La necesidad de reportar el método de la selección de los participantes y la asignación a los grupos experimentales debe ser obvia en este momento.

Dolinski y Nawrat: miedo-luego-alivio y sumisión

Los estudios sobre sumisión han representado gran interés para los psicólogos sociales. En el capítulo 17, donde se analizó la ética al realizar investigación en ciencias del comportamiento, se mencionó la influencia del estudio de Milgrim sobre la forma en que ahora se lleva a cabo la investigación. Milgrim, si se recuerda, se interesó en saber por qué durante la Segunda Guerra Mundial los nazis obedecían órdenes y cometían actos inenarrables de brutalidad hacia otros seres humanos. En un estudio de Dolinski y Nawrat (1998), se exploró otro método para inducir a la sumisión. Éste fue un método utilizado por los nazis y los estalinistas para obligar a los prisioneros polacos a testificar en contra de

sí mismos, sus amigos o sus familias. Dolinski y Nawrat llamaron a este método "miedo-luego-alivio", el cual implica poner a un prisionero en un estado de gran ansiedad por medio de gritos y amenazas de los carceleros hacia él. Después de alcanzar el nivel de miedo deseado, los estímulos generadores de ansiedad se eliminan abruptamente; entonces, se trata amablemente al prisionero. El resultado común de este procedimiento es la intensificación de la conducta de sumisión. Dolinski y Nawrat afirman que la sumisión se debe a la reducción del miedo y no al miedo en sí. Aunque Dolinski y Nawrat utilizan un ejemplo muy extremo para ilustrar su idea, ellos también explican que el método con frecuencia se utiliza de alguna manera y forma por diadas en la vida diaria. Puede ocurrir entre padre e hijo, maestro y estudiante, y entre empleado y patrono. La policía usa tácticas similares con su rutina del "policía bueno-policía malo", que por lo común incluye a un oficial de policía ("policía malo") que regaña, grita y amenaza a un sospechoso. Cuando el sospechoso alcanza un alto nivel de ansiedad, otro oficial de policía ("policía bueno") quita al "policía malo" y le habla amable y dulcemente al prisionero. Los terroristas también utilizan este método con los rehenes.

Dolinski y Nawrat diseñaron y realizaron cuatro experimentos para probar la eficacia del método de "miedo-luego-alivio" para inducir la sumisión. Aquí se describirá uno de estos experimentos, en el cual 120 estudiantes de preparatoria voluntarios de Opole, Polonia, fueron asignados aleatoriamente a una de tres condiciones experimentales. A todos los participantes se les avisó que participarían en un estudio sobre los efectos del castigo en el aprendizaje. El grupo 1 experimentó ansiedad; se les indicó que les sería aplicado un choque eléctrico leve, no doloroso, por cada error que cometieran. Los participantes del grupo 2 experimentaron ansiedad que después fue reducida. Al inicio se les dio la misma explicación que al grupo 1, pero después se les dijo que participarían en un estudio diferente, el cual implicaba coordinación visomotora y no incluía choques eléctricos. El grupo 3 era la condición control. A estos participantes se les indicó que iban a participar en un estudio sobre coordinación visomotora. Durante el periodo de espera, antes del inicio del experimento, se le pidió a cada participante contestar un cuestionario sobre ansiedad. Después de completar el cuestionario, una estudiante cómplice del experimentador, pero que parecía estar totalmente desligada del experimento, se presentó y le pidió a cada participante unirse a una acción de caridad para un orfanatorio. A quienes obedecieron o aceptaron se les preguntó cuántas horas estaban dispuestos a trabajar para dicha causa.

La variable independiente manipulada en este estudio fue el nivel de ansiedad inducida y de alivio. Las variables dependientes fueron la sumisión, la cantidad de ansiedad y el número de horas donadas para una buena causa. Con el uso de un análisis de varianza de un factor, Dolinski y Nawrat obtuvieron un valor *F* significativo. El grupo 2, el cual experimentó ansiedad que se redujo después, tuvo la tasa más alta de sumisión y el mayor

TABLA 21.1 *Niveles de ansiedad, sumisión, número de días dispuestos para ser voluntario por ansiedad inducida y valores F (estudio de Dolinski y Nawrat)*

Condición del grupo	Ansiedad media reportada	Porcentaje de sumisión	Número medio de días como voluntario
Estudio de choque eléctrico	53.25	37.5	0.625
Estudio de choque eléctrico cambiado por estudio de coordinación visomotora	43.05	75.0	1.150
Estudio de coordinación visomotora	34.45	52.5	1.025
Valor *F*	108.9 ($p < .00001$)	6.13 ($p < .003$)	2.11 ($p > .05$)

número de días dispuestos para la caridad. El nivel de ansiedad de cada grupo resultó en la dirección esperada. El grupo 1 experimentó el mayor nivel de ansiedad, seguido por el grupo 2 y luego el grupo 3. La tabla 21.1 presenta el resumen de los datos del estudio.

Los resultados del estudio apoyaron la hipótesis de Dolinski y Nawrat de que fue el "miedo-luego-alivio", y no la emoción del miedo en sí, lo que llevó a un nivel de sumisión mayor. Crear tan sólo un estado de ansiedad en las personas no es suficiente para inducir sumisión. De hecho, tal estudio encontró que los participantes del grupo 1 (ansiedad inducida), quienes sintieron la mayor cantidad de angustia, se sometieron en menor medida que los participantes del grupo 3 (control-ansiedad baja o sin ansiedad).

Diseños factoriales

El diseño básico general continúa siendo el diseño 20.1, aunque la variación del patrón básico de grupo experimental-grupo control se altera drásticamente al agregar otros factores experimentales o variables independientes. Siguiendo una definición previa del análisis factorial de varianza, *el diseño factorial es la estructura de la investigación, en la cual se yuxtaponen dos o más variables independientes para estudiar sus efectos independientes e interactivos sobre una variable dependiente.*

Al inicio, el lector puede encontrar un poco difícil ajustar la estructura factorial dentro del paradigma general de grupo experimental-grupo control del diseño 20.1. Sin embargo, el análisis de la generalización de la noción de grupo control en el capítulo 20 quizás aclaró las relaciones entre el diseño 20.1 y los diseños factoriales. La discusión continúa ahora. Se tienen las variables independientes A y B, y la variable dependiente Y. El diseño factorial más simple, el de 2×2, tiene tres posibilidades: tanto A como B son variables activas; A es activa, B es atributiva (o a la inversa); y tanto A como B son variables atributo. (La última posibilidad, ambas variables independientes son atributo, es el caso no experimental. No obstante, como se indicó anteriormente, tal vez no sea recomendable utilizar el análisis de varianza con variables independientes no experimentales.) Como siempre, A puede dividirse en A_1 y A_2, condición experimental y control, con la variable independiente adicional B dividida en B_1 y B_2. Puesto que esta estructura ahora resulta familiar, sólo es necesario discutir uno o dos detalles del procedimiento.

El procedimiento ideal de asignación de participantes consiste en asignarlos aleatoriamente a las cuatro casillas. Si tanto A como B son variables activas, esto se vuelve factible y fácil; tan sólo se les dan a los participantes números arbitrarios de 1 a N (donde N es el número total de participantes). Después, con una tabla de números aleatorios, se anotan números de 1 a N conforme aparecen en la tabla, los cuales se colocan en cuatro grupos conforme aparecen y después se asignan los cuatro grupos de participantes a las cuatro casillas. Para estar seguros, los grupos de participantes también se asignan aleatoriamente a los tratamientos experimentales (las cuatro casillas). Los grupos se denominan como 1, 2, 3 y 4, y después se extraen estos números de una tabla de números aleatorios. Suponga que la tabla produce los números en este orden: 3, 4, 1 y 2; entonces se asigna a los participantes del grupo 3 a la casilla superior izquierda; a los participantes del grupo 4, a la casilla superior derecha, etcétera.

Con frecuencia B será una variable atributo como género, inteligencia, rendimiento, ansiedad, autopercepción o raza. La asignación de los participantes debe verse alterada. Primero, puesto que B es una variable atributo, no existe posibilidad de asignar a los participantes a B_1 y B_2 de manera aleatoria. Si B es la variable género, lo mejor que se puede hacer es asignar aleatoriamente primero a los hombres a las casillas A_1B_1 y A_2B_1, y luego a las mujeres a las casillas A_1B_2 y A_2B_2.

Diseños factoriales con más de dos variables

Con frecuencia es posible mejorar el diseño y aumentar la información obtenida de un estudio añadiendo grupos. En lugar de A_1 y A_2 y de B_1 y B_2, un experimento podría beneficiarse al utilizar A_1, A_2, A_3 y A_4; y B_1, B_2 y B_3. Los problemas prácticos y estadísticos se incrementan y algunas ocasiones se tornan bastante difíciles conforme se agregan variables. Suponga que se tiene un diseño de $3 \times 2 \times 2$ que tiene $3 \times 2 \times 2 = 12$ celdillas, cada una de las cuales debe tener por lo menos dos participantes, y de preferencia muchos más. (Es posible, pero no muy sensible, incluir solamente un participante por casilla si es posible tener más. Por supuesto que existen diseños con sólo un participante por casilla. Esto se estudia en el capítulo 22.) Si se decide que son necesarios 10 participantes por casilla, entonces deberán obtenerse y asignarse aleatoriamente $12 \times 10 = 120$ participantes. El problema se vuelve más complicado con una variable más, y también resulta más difícil la manipulación práctica de la situación de investigación. Sin embargo, la manipulación exitosa de dicho experimento permite probar varias hipótesis y brinda una gran cantidad de información. Las combinaciones de diseños de tres, cuatro y cinco variables ofrecen una amplia variedad de diseños posibles: $2 \times 5 \times 3$, $4 \times 4 \times 2$, $3 \times 2 \times 4 \times 2$, $4 \times 3 \times 2 \times 2$, etcétera.

Ejemplos de investigación de diseños factoriales

En el capítulo 14 se describieron ejemplos de diseños factoriales de dos y tres dimensiones. (Se recomienda un repaso de estos ejemplos, pues el razonamiento que subyace al diseño esencial ahora puede captarse con mayor facilidad.) Como en el capítulo 14 se presentó un número suficiente de ejemplos de diseños factoriales, los ejemplos mostrados aquí se dedican a estudios con características poco usuales o resultados muy interesantes.

Sigall y Ostrove: atractivo y crimen

A menudo se afirma que a las mujeres atractivas se les trata de forma diferente que a los hombres o que a las mujeres poco atractivas. En la mayoría de los casos, quizá, las reacciones son "favorables": las mujeres atractivas tal vez tienen una mayor probabilidad, que las mujeres no atractivas, de recibir la atención y los favores del mundo. No obstante, ¿será posible que su atractivo sea desventajoso en algunas situaciones? Sigall y Ostrove (1975) plantearon la pregunta: ¿cómo se relaciona el atractivo físico de un criminal acusado con las sentencias judiciales, y cómo la naturaleza del crimen interactúa con el atractivo? Ellos pidieron a sus participantes que asignaran sentencias, en años, a delitos de estafa y robo de acusadas atractivas, no atractivas y controles. En la tabla 21.2 se presenta el paradigma factorial del experimento, junto con los resultados. (Se evitó la descripción de muchos detalles experimentales; éstos fueron bien manejados.)

TABLA 21.2 *Sentencias medias en años de acusadas atractivas, no atractivas y control, por estafa y robo (estudio de Sigall y Ostrove)*[a]

	Condición de la acusada		
	Atractiva	**No atractiva**	**Control**
Estafa	5.45	4.35	4.35
Robo	2.80	5.20	5.10

[a] $N = 120$, 20 por casilla; F (interacción) = 4.55 ($p < .025$).

En el caso de robo, la acusada robó $2 200 en un rascacielos. En la situación de estafa, la acusada se congració con un soltero de mediana edad y lo estafó con $2 200. Observe que las condiciones de no atractiva y control no difirieron mucho entre sí. Tanto la situación *atractiva-estafa* (5.45) como la situación *atractiva-robo* (2.80) difirieron de las otras dos condiciones, ¡pero en direcciones opuestas! La situación *atractiva-estafa* recibió la mayor sentencia media: 5.45 años; mientras que la situación *atractiva-robo* recibió la menor sentencia media: 2.80 años. Los estadísticos apoyan el resumen verbal previo —la interacción fue estadísticamente significativa: la F de *atractiva-delito*, con 2 y 106 grados de libertad, fue de 4.55, $p < .025$; es decir, las acusadas atractivas tienen una ventaja sobre las acusadas no atractivas, excepto cuando sus crímenes están relacionados con su atractivo (estafa)—.

Quilici y Mayer: ejemplos, esquema y aprendizaje

¿Ayudan los ejemplos a que los estudiantes aprendan estadística? Ésta fue la pregunta básica planteada por los científicos cognitivos Quilici y Mayer (1966). En su estudio sobre la solución analítica de problemas, Quilici y Mayer examinaron sólo uno de tres procesos que definen el pensamiento analógico. Ellos se interesaron tan sólo en el proceso de reconocimiento que implica dos técnicas: 1) enfoque en las similitudes superficiales entre el ejemplo y el problema real que debe resolverse, o 2) enfoque en las similitudes estructurales.

Las similitudes superficiales tratan con los atributos compartidos de objetos en la historia del problema. En la similitud estructural lo importante son las relaciones compartidas entre objetos, tanto en el ejemplo como en el problema. Para estudiar dicho fenómeno, Quilici y Mayer utilizaron el aprendizaje de la resolución de problemas escritos en estadística. Tuvieron la impresión de que los estudiantes que aprenden la estructura de problemas estadísticos escritos serían más capaces de resolver otros problemas que enfrentaran en el futuro, al clasificarlos apropiadamente en el método de análisis estadístico correcto (por ejemplo, prueba *t*, correlación, etcétera). A continuación se presentan cuatro ejemplos para ilustrar las diferencias entre las similitudes superficiales y estructurales.

Ejemplo 1

Un experto en personal desea determinar si los mecanógrafos con experiencia son capaces de teclear más rápido que los mecanógrafos sin experiencia. A 20 mecanógrafos con experiencia y a 20 sin experiencia se les aplica una prueba de mecanografía. Se registra el número promedio de palabras tecleadas por minuto de cada uno de ellos.

Ejemplo 2

Un experto en personal desea determinar si la experiencia en mecanografía se relaciona con velocidades más rápidas al teclear. Se les pide a 40 mecanógrafos que informen cuántos años han trabajado como tales y se les aplica una prueba de mecanografía para determinar su número promedio de palabras tecleadas por minuto.

Ejemplo 3

Después de revisar los datos sobre el clima de los últimos 50 años, una meteoróloga afirma que la precipitación anual varía con la temperatura promedio. Para cada uno de los 50 años ella verifica la caída de lluvia anual y la temperatura promedio.

Ejemplo 4

Un decano universitario afirma que los lectores eficientes obtienen mejores calificaciones que los lectores ineficientes. Se registran las calificaciones promedio de 50

estudiantes de primer año, quienes obtuvieron una alta puntuación en una prueba de lectura de comprensión; y de 50 estudiantes de primer año que obtuvieron una baja puntuación en la misma prueba.

Si se examinan estos cuatro problemas tomados de Quilici y Mayer (1996, p. 146), el ejemplo 1 y el ejemplo 2 tendrían las mismas características superficiales; ambos tratan de mecanógrafos y de tecleo. Para resolver el ejemplo 1 se utilizaría una prueba *t*, para comparar a los mecanógrafos con experiencia con aquellos sin experiencia. No obstante, para resolver el ejemplo 2 se utilizaría una correlación, ya que la cuestión requiere de una relación entre la experiencia en mecanografía y el número promedio de palabras tecleadas por minuto. Por lo tanto, los ejemplos 1 y 2 serían estructuralmente diferentes. El ejemplo 3 también analiza la relación entre dos variables: cantidad de lluvia y temperatura. Este ejemplo tendría la misma estructura del ejemplo 2, pero una superficie diferente. Poseen la misma estructura, ya que ambos requieren del uso de una correlación para resolver el problema. El ejemplo 4 y el ejemplo 1 tienen la misma estructura; pero una superficie diferente.

Quilici y Mayer diseñaron un estudio para determinar si la experiencia con ejemplos fomentaba la construcción de esquemas estructurales. Supusieron que la exposición a problemas estadísticos escritos haría a los estudiantes más sensibles a las características estructurales que a las superficiales, en futuros problemas escritos. Los estudiantes que no son expuestos a ejemplos estadísticos escritos no exhibirían dicha conducta. Ellos también hipotetizaron que aquellos a quienes se expusiera a tres ejemplos serían capaces de exhibir la conducta en un mayor grado que aquellos que fueran expuestos a sólo un ejemplo. Estos investigadores utilizaron un diseño factorial de 3 × 2. La primera variable independiente fueron las características estructurales (prueba *t*, chi cuadrada y correlación). La segunda variable independiente fueron las características superficiales (mecanografía, clima, fatiga mental y lectura). Hubo dos variables dependientes: una puntuación de uso estructural y una puntuación de uso superficial. Los participantes fueron asignados aleatoriamente a las condiciones de tratamiento. Un análisis de varianza de dos factores confirmó su hipótesis de que aquellos expuestos a ejemplos utilizarían un esquema de base estructural; mientras que quienes no fueron expuestos a ejemplos no lo harían. Sin embargo, no hubo una diferencia estadística entre aquellos expuestos a tres ejemplos y quienes recibieron un ejemplo.

Hoyt: Conocimiento del maestro y rendimiento del alumno

Ahora se describe un estudio educativo realizado hace muchos años, que se planeó para responder una importante pregunta teórica y práctica, que ilustra claramente un diseño factorial complejo. La pregunta de investigación fue: ¿cuáles son los efectos sobre el rendimiento y las actitudes de los alumnos, si a los maestros se les informa respecto a las características de sus alumnos? El estudio de Hoyt (1955) exploró diversos aspectos de la pregunta básica y utilizó un diseño factorial para incrementar la validez tanto interna como externa de la investigación. El primer diseño se utilizó tres veces para cada uno de los tres participantes escolares; y el segundo y el tercero se usaron dos veces, una vez en cada uno de los dos sistemas escolares.

El paradigma del primer diseño se presenta en la figura 21.1. Las variables independientes fueron los tratamientos, la habilidad, el sexo y las escuelas. Los tres tratamientos autoexplicativos fueron: sin información (*N*), puntuaciones de la prueba (*P*) y puntuaciones de la prueba más otra información (*PO*). Los niveles de habilidad eran CI *alto*, *medio* y *bajo*. Las variables *género* y *escuelas* resultan obvias. Los estudiantes de octavo grado fueron asignados aleatoriamente en cuanto al género y niveles de habilidad. Ayudará a compren-

FIGURA 21.1

		N		*P*		*PO*	
		Hombre	**Mujer**	**Hombre**	**Mujer**	**Hombre**	**Mujer**
Escuela *A*	CI alto						
	CI medio						
	CI bajo						
Escuela *B*	CI alto						
	CI medio						
	CI bajo						

Medidas de la variable dependiente

der el diseño si se examina la forma de una tabla final de análisis de varianza del diseño. Sin embargo, antes de hacerlo debe notarse que los resultados de *rendimiento* en su mayoría fueron indeterminantes (o negativos). Las razones *F*, con una excepción, no fueron significativas. Las actitudes de los alumnos hacia los profesores, por otro lado, parecieron mejorar cuando los maestros incrementaron su conocimiento sobre los alumnos: un hallazgo interesante y potencialmente importante. La tabla del análisis de varianza se presenta en la tabla 21.3. ¡Un experimento que produce 14 pruebas! En efecto, algunas de estas pruebas no son importantes y pueden ignorarse. Las pruebas de mayor importancia (marcadas con asteriscos en la tabla) son aquellas que incluyen la variable tratamiento. La prueba más importante es entre tratamientos, el primero de los efectos principales. Quizá

TABLA 21.3 *Fuentes de varianza y grados de libertad para un diseño factorial de 3 × 3 × 2 × 2, con las variables tratamientos, habilidad, sexo y escuela (se omitieron los grados de libertad totales y dentro)*

Fuente	*gl*
Efectos principales	
Entre tratamientos*	2
Entre niveles de habilidad	2
Entre géneros	1
Entre escuelas	1
Interacciones de primer orden	
Interacción: tratamientos × habilidad	4
Interacción: tratamientos × género*	2
Interacción: tratamientos × escuela*	2
Interacción: habilidad × género	2
Interacción: habilidad × escuela	2
Interacción: género × escuela	1
Interacciones de segundo orden	
Interacción: tratamientos × habilidad × género*	4
Interacción: tratamientos × habilidad × escuela	4
Interacción: habilidad × género × escuela	2
Interacciones de tercer orden	
Interacción: tratamientos × habilidad × género × escuela	4

de igual importancia sean las interacciones que incluyen tratamientos. Tome la interacción tratamiento × sexo; si resulta significativa, entonces quiere decir que la cantidad de información que un maestro posee sobre los estudiantes ejerce una influencia en el rendimiento de estos últimos; pero los niños se ven influenciados de manera diferente que las niñas. A los niños con maestros que tienen información sobre sus alumnos les puede ir mejor que a los niños cuyos maestros no poseen dicha información; mientras que puede suceder lo contrario con las niñas, o puede no resultar ninguna diferencia en uno u otro sentido.

Las interacciones de segundo orden o triples son más difíciles de interpretar. Parece que rara vez son significativas; sin embargo, requieren un estudio especial. Las tablas de tabulación cruzadas de las medias quizás sean la mejor opción; aunque los métodos gráficos, como se analizó previamente, a menudo son ilustrativos. El lector encontrará una guía en el libro de Edward (1984) y en el manuscrito de Simon (1976).

Evaluación de los diseños de sujetos aleatorizados

Todos los diseños de sujetos aleatorizados son variantes o extensiones del diseño 20.1, el diseño básico de grupo experimental-grupo control, en el cual los participantes son asignados de forma aleatoria a los grupos experimental y control. De esta manera, incluyen las fortalezas del diseño básico, de las cuales la más importante es la característica de aleatorización y la consecuente habilidad para suponer la igualdad aproximada preexperimental de los grupos experimentales, en todas las variables independientes posibles. Se controlan la historia y la maduración ya que pasa muy poco tiempo entre la manipulación de X, y la observación y la medida de Y. No existe la posibilidad de contaminación debida al pretest.

Las otras fortalezas de estos diseños, que surgen de las múltiples variaciones posibles, son la flexibilidad y la aplicabilidad, las cuales sirven para ayudar a resolver muchos problemas en la investigación del comportamiento, puesto que parecen ajustarse particularmente bien a los tipos de problemas del diseño que surgen de problemas e hipótesis científicos, tanto sociales como educativos. El diseño de un factor, por ejemplo, incorpora cualquier número de métodos y la comprobación de métodos es una necesidad educativa importante. Las variables que constantemente necesitan control en la investigación del comportamiento —género, inteligencia, aptitud, clase social, escuelas y muchas otras— pueden incorporarse a los diseños factoriales, y así se controlan. También, con los diseños factoriales es posible realizar mezclas de variables activas y atributo —otra importante necesidad—. Sin embargo, también existen debilidades.

Una crítica consiste en que los diseños de sujetos aleatorizados no permiten pruebas de la igualdad de los grupos, como lo hacen los diseños antes-después (pretest-postest). En realidad ésta no es una crítica válida, por dos razones: 1) como se ha visto, con suficientes participantes y aleatorización se supone que los grupos son iguales; y 2) es posible verificar la igualdad de los grupos en variables diferentes de Y, la variable dependiente. Para la investigación educativa, en los expedientes escolares existe información sobre inteligencia, aptitud y rendimiento, por ejemplo. Datos pertinentes para investigación en sociología y en ciencias políticas a menudo están disponibles en expedientes de condados y de distritos electorales.

Otra debilidad es de tipo estadístico. Debe tenerse igual número de casos en las casillas de los diseños factoriales. Es posible trabajar con n desiguales; pero es insensato y representa una amenaza para la interpretación. La eliminación aleatoria de casos o el uso de métodos de casos faltantes contrarrestan pequeñas discrepancias (véase Dear, 1959;

Gleason y Staelin, 1975, dos excelentes referencias sobre la estimación de datos faltantes). Lo anterior impone una limitación sobre el uso de dichos diseños, ya que no siempre es posible tener números iguales en cada casilla. Los diseños aleatorizados de un factor no son tan delicados: los números desiguales no representan un problema difícil. Cómo ajustar y analizar datos con n desiguales constituye un problema complejo, polémico y muy discutido. Para revisar una discusión en contexto, principalmente del análisis de varianza, se recomienda consultar Snedecor y Cochran (1989). Respecto a la discusión en el contexto de la regresión múltiple, la cual representa una mejor solución del problema, véase Kerlinger y Pedhazur (1973) y Pedhazur (1996). Las discusiones de Pedhazur son detalladas y con autoridad; revisa los temas y sugiere soluciones.

En comparación con los diseños de grupos apareados, los diseños de sujetos aleatorizados por lo común son menos precisos, es decir, el término del error normalmente es mayor, si lo demás permanece igual. No obstante, es dudoso que ello sea un motivo de preocupación. En ciertos casos, en efecto lo es —por ejemplo, cuando se requiere de una prueba muy sensible para una hipótesis—. Sin embargo, en gran parte de la investigación del comportamiento tal vez sea deseable considerar como no significativo cualquier efecto que sea insuficientemente poderoso para hacerse sentir sobre y por arriba del ruido aleatorio de un diseño de sujetos aleatorizados.

De cualquier forma, entonces, éstos son diseños poderosos, flexibles, útiles y de amplia aplicación. En la opinión de los autores, son los mejores diseños disponibles, quizá los primeros a considerarse al planear el diseño de un estudio de investigación.

Grupos correlacionados

Existe un principio básico detrás de todos los diseños de grupos correlacionados: hay varianza sistemática en las medidas de la variable dependiente, debida a la correlación entre los grupos *en alguna variable relacionada con la variable dependiente*. Esta correlación y su varianza concomitante puede introducirse en las medidas y en el diseño de tres formas:

1. empleando las mismas unidades, por ejemplo participantes, en cada uno de los grupos experimentales,
2. apareando las unidades respecto a una o más variables independientes que estén relacionadas con la variable dependiente, y
3. usando más de un grupo de unidades en el diseño, como clases o escuelas.

A pesar de las aparentes diferencias entre las tres formas para introducir la correlación en las medidas de la variable dependiente, básicamente son las mismas. Ahora se examinarán las implicaciones que tiene este principio básico para el diseño, y se analizarán las formas de implementarlo.

El paradigma general

Con excepción de los diseños factoriales correlacionados y de los llamados diseños anidados, todos los paradigmas del análisis de varianza de diseños de grupos correlacionados se bosquejan fácilmente. El término *grupo* debería utilizarse para indicar conjuntos de puntuaciones; así no hay confusión cuando un experimento de ensayos repetidos se clasifica como un diseño multigrupal. El paradigma general se presenta en la figura 21.2. Para enfatizar las fuentes de varianza, se indican las medias de las columnas y de los renglones; también se incluyen las medidas individuales de la variable dependiente (Y).

Resulta útil conocer el sistema de subíndices de los símbolos utilizados en matemáticas y estadística. Una tabla rectangular de números se llama *matriz*. Los elementos de una matriz son letras y/o números. Cuando se utilizan letras, es común identificar cualquier elemento particular de la matriz con dos (en ocasiones más) subíndices. El primero de ellos indica el número de la posición del renglón; y el segundo, el número de la posición de la columna. Y_{32}, por ejemplo, indica la medida de Y en el tercer renglón y en la segunda columna; Y_{52} indica la medida de Y del quinto renglón y de la segunda columna. También se acostumbra generalizar tal sistema al añadir subíndices de letras. En este libro i simboliza cualquier número de renglón, y j cualquier número de columna. Cualquier número de la matriz se representa por Y_{ij}. Cualquier número del tercer renglón es Y_{3j}; y cualquier número de la segunda columna es Y_{i2}.

Puede notarse que existen dos fuentes de varianza sistemática: aquella debida a las columnas o los tratamientos, y aquella debida a los renglones (diferencias individuales o de unidad). El análisis de varianza debe ser de dos factores.

El lector que ya estudió el análisis sobre la varianza de correlación del capítulo 15, donde se presentaron los estadísticos y algunos de los problemas de los diseños de grupos correlacionados, no tendrá dificultad con el razonamiento de la varianza de la figura 21.2. La intención del diseño consiste en maximizar la varianza entre tratamientos, identificar la varianza entre unidades y la varianza del error (residual). El principio del maxmincon aplica aquí como en cualquier otro lado. La única diferencia, en realidad, entre los diseños de grupos correlacionados y los sujetos aleatorizados es la varianza de los renglones o de las unidades.

Unidades

Las unidades utilizadas no alteran el principio de la varianza. El término *unidad* se usa deliberadamente para enfatizar que las unidades pueden ser personas o participantes, clases, escuelas, distritos, ciudades e incluso naciones. En otras palabras, la "unidad" es un rubro generalizado que puede representar muchos tipos de entidades. La consideración importante es si las unidades —cualesquiera que sean— difieren o no entre sí; si difieren, entonces se introduce *varianza entre unidades*. En este sentido, hablar de grupos o participantes correlacionados es lo mismo que hablar de varianza entre grupos o participantes. El concepto de diferencias individuales se extiende a *diferencias* de unidades.

El valor real del diseño de grupos correlacionados, más allá de permitir al investigador aislar y estimar la varianza debida a la correlación, es la guía que permite al investigador diseñar investigaciones para aprovechar las diferencias que frecuentemente existen

FIGURA 21.2

	Tratamientos							
Unidades	X_1	X_2	X_3	.	.	.	X_k	**Renglones**
1	Y_{11}	Y_{12}	Y_{13}	.	.	.	Y_{1k}	M_1
2	Y_{21}	Y_{22}	Y_{23}	.	.	.	Y_{2k}	M_2
	Y_{31}	Y_{32}	Y_{33}	.	.	.	Y_{3k}	M_3
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
n	Y_{n1}	Y_{n2}	Y_{n3}	.	.	.	Y_{nk}	M_n
	M_{c1}	M_{c2}	M_{c3}	.	.	.	M_{ck}	(M_t)

entre las unidades. Si un estudio de investigación incluye diferentes clases de la misma escuela, éstas son una posible fuente de varianza; por lo tanto, sería sensato utilizar las "clases" como unidades en el diseño. Las diferencias bien conocidas entre escuelas son fuentes de varianza muy importantes en la investigación del comportamiento; pueden manejarse como un diseño factorial o en la forma de los diseños de este capítulo. De hecho, si se observa con detenimiento un diseño factorial con dos variables independientes, una como *escuelas*, y un diseño de grupos correlacionados con las unidades *escuelas*, en esencia se encuentra el mismo diseño. Estudie la figura 21.3; a la izquierda hay un diseño factorial y a la derecha un diseño de grupos correlacionados; sin embargo, ¡se ven iguales! Lo son respecto al principio de la varianza. (La única diferencia sería el número de puntuaciones en las casillas y el tratamiento estadístico.)

Diseño de un grupo con ensayos repetidos

En el diseño de un grupo con ensayos repetidos, como su nombre lo indica, a un grupo se le dan diferentes tratamientos en diferentes momentos. En un experimento sobre aprendizaje, el mismo grupo de participantes puede recibir varias tareas de complejidad diferente, o la manipulación experimental tal vez sea la presentación de principios de aprendizaje en órdenes distintos, por ejemplo, de simple a complejo, de complejo a simple, del todo a la parte, de la parte al todo.

Anteriormente se indicó que el mejor apareamiento posible de los participantes consiste en aparearlos consigo mismos. Las dificultades del empleo de tal solución del problema de control también se mencionó. Una de ellas se refiere a la sensibilización del pretest, el cual puede producir una interacción entre el pretest y la variable manipulada experimentalmente. Otra dificultad reside en que los participantes maduren y aprendan a través del tiempo. Un participante que ha experimentado uno o dos ensayos de una manipulación experimental y que enfrenta un tercer ensayo, ahora es una persona diferente de la que enfrentó el primer ensayo. Las situaciones experimentales difieren mucho, por supuesto. En algunas situaciones los ensayos repetidos quizá no afecten en exceso el desempeño de los participantes en ensayos posteriores; en otras situaciones posiblemente sí. El problema sobre cómo aprenden los individuos, o cómo se sensibilizan excesivamente durante un experimento, es difícil de resolver. En resumen, la *historia*, la *maduración* y la *sensibilización* son posibles debilidades de los ensayos repetidos. El efecto de regresión también es una debilidad, ya que, como se vio en un capítulo previo, los individuos con bajas puntuaciones tienden a obtener puntuaciones más altas; y los individuos con altas puntuaciones tienden a obtener puntuaciones más bajas en el postest, debido simplemente a la correlación imperfecta entre los grupos. Por supuesto, se requiere de un grupo control.

A pesar de las dificultades básicas de tiempo, habrá ocasiones en que un diseño de un grupo con ensayos repetidos sea útil. En efecto, en el análisis de datos de "tiempo", éste es

FIGURA 21.3

	Tratamientos			Tratamientos	
Escuelas	A_1	A_2	Escuelas	A_1	A_2
B_1			1		
B_2			2		
B_3			3		
Diseño factorial			Diseño de grupos correlacionados		

el diseño implícito. Si se tienen series de mediciones del crecimiento en niños, por ejemplo, los distintos momentos en que se hicieron las mediciones corresponden a los tratamientos. El paradigma del diseño es el mismo mostrado en la figura 21.2. Tan sólo se sustituyen "participantes" por "unidades" y se anotan $X_1, X_2, \ldots$ como "ensayos".

A partir de este paradigma general, es posible derivar casos especiales. El caso más simple es el diseño pretest-postest de un grupo, diseño 19.2(a), donde se aplicó un tratamiento experimental a un grupo de participantes, precedido de un pretest y seguido de un postest. Puesto que las debilidades de tal diseño ya se mencionaron, no es necesario ampliar la discusión. No obstante, debe notarse que este diseño, especialmente en su forma no experimental, se aproxima bastante a muchas observaciones y pensamientos de sentido común. Una persona observa prácticas educativas hoy y decide que no son buenas. Para realizar dicho juicio, uno compara implícita o explícitamente las prácticas educativas de hoy con las del pasado. De un posible número de causas, dependiendo del sesgo particular, el investigador seleccionará una o más razones por las que él considera lamentable el estado de los asuntos educativos: "educación progresiva", "educacionistas", "degeneración moral", "carencia de principios religiosos firmes", etcétera.

Diseños de dos grupos: grupo experimental-grupo control

Se trata de un diseño con dos formas, la mejor de las cuales (repetida aquí) se describió en el capítulo 20 como diseño 20.2:

$[Ap_a]$	X	Y	(Experimental)
	$\sim X$	Y	(Control)

En este diseño primero se aparea a los participantes y después se les asigna aleatoriamente a los grupos experimental y control. En la otra forma, se aparea a los participantes, pero no se les asigna a los grupos experimental y control de manera aleatoria. El último diseño se indica simplemente por medio de la eliminación del subíndice a (asignación aleatoria) de Ap_a, que indica el apareamiento de los sujetos y su asignación aleatoria a los grupos (descrito en el capítulo 19 como el diseño 19.4, uno de los diseños menos adecuados).

El paradigma estadístico de este caballo de batalla de los diseños se presenta en la figura 21.4. La inserción de los símbolos para las medias indica las dos fuentes de varianza

FIGURA 21.4

Pares	Tratamientos X_e	Tratamientos X_c	
1	Y_{1e}	Y_{1c}	M_1
2	Y_{2e}	Y_{2c}	M_2
3	Y_{3e}	Y_{3c}	M_3
.	.	.	.
.	.	.	.
n	Y_{ne}	Y_{nc}	M_n
	M_e	M_c	

sistemática: *tratamientos* y *pares*, columnas y renglones. Éste contrasta claramente con los diseños aleatorizados en una sección previa de este capítulo, donde la única varianza sistemática eran los *tratamientos* o columnas.

La variante más común del diseño de dos grupos, grupo experimental-grupo control es el diseño pretest-postest de dos grupos [véase diseño 20.3(*b*)]. El paradigma estadístico del diseño y su lógica se analizarán más adelante.

Ejemplos de investigación de los diseños de grupos correlacionados

Se han publicado cientos de estudios del tipo de grupos correlacionados. Los diseños usados con mayor frecuencia son los de participantes apareados, o los mismos participantes con pretest y postest. Sin embargo, los diseños de grupos correlacionados no se limitan a dos grupos; por ejemplo, a los mismos participantes se les puede aplicar más de dos tratamientos experimentales. Los estudios que se describen a continuación se eligieron no sólo porque ilustran los diseños de grupos correlacionados, el apareamiento y los problemas de control, sino también porque son importantes histórica, psicológica o educativamente.

Estudio de transferencia del aprendizaje de Thorndike

En 1924, E. L. Thorndike publicó un notable estudio sobre el supuesto efecto de ciertas materias en la inteligencia de los estudiantes. Los estudiantes fueron apareados de acuerdo con las puntuaciones en la forma A de la medida de la variable dependiente, la *inteligencia*. Esta prueba también sirvió como un pretest. La variable independiente fue un *estudio de un año de los participantes*, en materias tales como historia, matemáticas y latín. Al final del año se les aplicó un postest, la forma B de la prueba de inteligencia. Thorndike (1924) utilizó un recurso ingenioso para separar el efecto diferencial de cada materia escolar, al aparear en la forma A de la prueba de inteligencia a aquellos alumnos que estudiaron, por ejemplo, inglés, historia, geometría y *latín*, con los alumnos que estudiaron inglés, historia, geometría y *taller*. Así, para estos dos grupos, él comparó los efectos diferenciales de *latín* y *taller*. Los incrementos en las puntuaciones finales de inteligencia se consideraron como un efecto conjunto del crecimiento y de las materias académicas estudiadas.

A pesar de sus debilidades, fue un estudio colosal. Thorndike estaba consciente de la falta de controles adecuados, como lo revela en el siguiente párrafo sobre los efectos de la selección.

> La principal razón por la cual los buenos pensadores parecen superficialmente haberlo hecho así al tomar ciertos estudios escolares es que los buenos pensadores han tomado dichos estudios... Cuando los buenos pensadores estudiaron griego y latín, tales estudios parecieron hacer buenos pensadores. Ahora que los buenos pensadores estudian física y trigonometría, éstas parecen formar buenos pensadores. Si los alumnos más capaces debieran estudiar educación física y arte dramático, entonces estas materias parecerían formar buenos pensadores (p. 98).

Thorndike señaló el camino de la investigación educativa controlada, el cual conlleva la disminución de explicaciones metafísicas y dogmáticas en la educación. Su trabajo dio un golpe a la teoría de "la frotación de la navaja" del entrenamiento mental, aquella que semejaba la mente con una navaja que podía afilarse frotándola sobre sujetos "duros".

No es fácil evaluar un estudio como éste, cuya índole e ingenuidad son impresionantes. Sin embargo, uno se pregunta sobre la adecuación de la variable dependiente, inteli-

gencia o habilidad intelectual. ¿Las materias escolares estudiadas durante un año pueden tener un gran efecto sobre la inteligencia? Además, el estudio fue no experimental. Thorndike midió la inteligencia de los estudiantes y dejó que operaran las variables independientes, *materias escolares*. Por supuesto que no era posible realizar ninguna aleatorización. Como se mencionó antes, él estaba consciente de tal debilidad en el control de su estudio, el cual es todavía un clásico que merece respeto y estudio cuidadoso, a pesar de sus debilidades en cuanto a historia y selección (se controló la maduración).

Miller y DiCara: aprendizaje de funciones autónomas

En un capítulo anterior se presentaron los datos de un estudio, del notable conjunto de estudios sobre el aprendizaje de funciones autónomas realizado por Miller y sus colegas (Miller, 1971; Miller y DiCara, 1968). Tanto los expertos como los novatos consideran que no es posible aprender y controlar las respuestas del sistema nervioso autónomo. Es decir, que respuestas glandulares y viscerales —latido cardiaco, secreción de orina y presión sanguínea, por ejemplo— se suponían más allá del "control" del individuo. Miller creía lo contrario, pues experimentalmente demostró que tales respuestas están sujetas al aprendizaje instrumental. La parte crucial de este método consistió en recompensar las respuestas viscerales cuando ocurrían. En el estudio (los datos se citaron en un capítulo previo de este libro) se recompensaba a las ratas cuando incrementaban o disminuían la secreción de orina. Se asignaron aleatoriamente 14 ratas a dos grupos llamados "ratas con incremento" y "ratas con disminución". Las ratas del primer grupo fueron recompensadas con estimulación cerebral (la cual había resultado efectiva para *incrementar* la secreción de orina); mientras que las ratas del último grupo fueron recompensadas por *disminuir* la secreción de orina durante un periodo de "entrenamiento" de 220 ensayos, en aproximadamente tres horas.

Para mostrar parte de los paradigmas experimental y analítico de este experimento, los datos antes-después de los periodos de entrenamiento de las ratas con incremento y las ratas con disminución, se incluyen en la tabla 21.4 (tomados de la tabla 1 de Miller y DiCara). Las medidas en la tabla son mililitros de orina secretada por minuto por cada 100 gramos de peso corporal. Observe que todas son cantidades muy pequeñas. El diseño de la investigación es una variante del diseño 20.3(*a*)

[*A*]	Y_a	X	Y_d	(Experimental)
	Y_a	$\sim X$	Y_d	(Control)

La diferencia es que $\sim X$, lo que en el diseño significa ausencia de tratamiento experimental para el grupo control, ahora significa recompensa por decremento en la secreción de orina. Por lo tanto, se altera el análisis usual de las medidas antes-después de los dos grupos.

El análisis se comprende mejor si se analizan los datos de la tabla 21.4, de manera diferente a como lo hicieron Miller y DiCara. (Ellos utilizaron pruebas *t*.) Aquí se realizó un análisis de varianza de dos factores (medidas repetidas) de los datos de las ratas con incremento, *antes-después* y de los datos de las ratas con disminución, *antes-después*. Las medias *antes* y *después* del grupo con incremento fueron .017 y .028, y las del grupo con disminución fueron .020 y .006. La razón *F* del grupo con incremento fue 43.875 (gl = 1.6): la *F* de las ratas con disminución fue 46.624. Ambas fueron altamente significativas. Sin embargo, las dos medias *antes*, .017 y .020, no fueron significativamente diferentes. En este caso, la comparación de las medias *después* de los dos grupos, la comparación acostumbrada con este diseño, probablemente no sea apropiada debido a que una era para el incremento y la otra para la disminución de la secreción de orina.

TABLA 21.4 *Datos de secreción de orina, ratas con incremento y ratas con disminución, entrenamiento antes-después (estudio de Miller y DiCara)*

Ratas con incremento[a]				*Ratas con disminución*[b]			
Ratas	**Antes**	**Después**	**Σ**	**Ratas**	**Antes**	**Después**	**Σ**
1	.023	.030	.053	1	.018	.007	.025
2	.014	.019	.033	2	.015	.003	.018
3	.016	.029	.045	3	.012	.005	.017
4	.018	.030	.048	4	.015	.006	.021
5	.007	.016	.023	5	.030	.009	.039
6	.026	.044	.070	6	.027	.008	.035
7	.012	.026	.038	7	.020	.003	.023
Medias	.017	.028			.020	.006	.023

[a] Incremento, antes-después: $F = 43.875$ ($p < .001$); $\omega^2 = .357$. Las medidas en la tabla son mililitros por minuto por cada 100 gramos de peso.
[b] Decremento, antes-después: $F = 46.624$ ($p < .001$); $\omega^2 = .663$.

Este estudio, con sus manipulaciones experimentales altamente controladas y sus análisis de "controles", constituye un ejemplo de la conceptualización imaginativa y del análisis competente disciplinado. El análisis anterior es un ejemplo, pero los autores del estudio hicieron mucho más. Por ejemplo, para estar más seguros de que el reforzamiento afectó únicamente la secreción de orina, compararon la frecuencia cardiaca (latidos por minuto) *antes-después*, tanto de las ratas con incremento como con disminución. Las medias fueron 367 y 412 para las ratas con incremento; y 373 y 390 para las ratas con disminución. Ninguna de las diferencias fue estadísticamente significativa. Comparaciones similares de la presión sanguínea y de otras funciones corporales tampoco fueron significativas.

Se recomienda a los estudiantes estudiar este excelente ejemplo de investigación en laboratorio hasta que comprendan claramente qué se hizo y por qué, lo cual los ayudará a aprender más sobre experimentos controlados, diseño de investigación y análisis estadístico, que la mayoría de los ejercicios en libros de texto. ¡Es un logro espléndido!

Tipper, Eissenberg y Weaver: efectos de la práctica sobre la atención selectiva

Cuando se habla de atención selectiva, algunos podrán recordar el estudio clásico de Stroop (1935), quien demostró el papel de la interferencia sobre la atención selectiva. Los estímulos irrelevantes compiten con los relevantes para lograr el control de la acción perceptual. Para aquellos que no estén familiarizados con dicho estudio, una parte memorable fue presentar a los participantes palabras como *verde* y *azul* impresas en rojo y amarillo. Luego se les pidió nombrar los colores en los que estaban escritas las palabras, pero en lugar de eso, los sujetos leían las palabras. Las personas tienen dificultad para suprimir el hábito de leer palabras aun cuando se les pide que no lo hagan. Para realizar dicha tarea de forma correcta, el participante debe concentrarse y evitar de manera consciente leer las palabras. Esta interferencia fue llamada el efecto Stroop. Desde la realización del famoso estudio de Stroop, se ha llevado a cabo un gran número de estudios sobre atención selectiva; el estudio de Tipper, Eissenberg y Weaver (1992) es uno de ellos. Este estudio es diferente, ya que discute varios aspectos respecto a numerosos estudios realizados sobre atención selectiva. Primero, Tipper y sus colaboradores hipotetizaron que cualquier experimento sobre atención selectiva que utilice participantes durante una hora o más puede estar conectando diferentes mecanismos perceptuales de los que se usan en la vida diaria. Los experimentos

de laboratorio casi siempre requieren que los participantes estén presentes durante aproximadamente una hora. En el periodo de una hora la experiencia experimental completa es aún novedosa. Es probable que la selectividad de atención se logre por medio de diferentes mecanismos conforme los estímulos se vuelven más familiares.

Tipper y sus colegas diseñaron un estudio para probar su hipótesis sobre atención selectiva utilizando un diseño dentro de sujetos. Todos los participantes experimentaron todas las condiciones del tratamiento. Ellos observaron el efecto de la interferencia sobre el tiempo de reacción y los errores. Hicieron que cada participante experimentara ambos niveles de interferencia: preparación negativa e inhibición de respuesta a través de 11 bloques o ensayos tomados durante cuatro días (efecto de la práctica). Sus resultados demostraron que hubo un efecto de interferencia ($F = 35.15$, $p < .001$) cuando se utilizó el tiempo de reacción como variable dependiente. Los tiempos de reacción fueron más largos cuando la distracción estuvo presente. También encontraron un efecto por la práctica (bloques) ($F = 9.62$, $p < .0001$) y ningún efecto de interacción. El efecto de la práctica indicó que la reacción de los participantes se torna más rápida con el incremento de la práctica. El hecho de que el efecto de interacción no fuese significativo indica que los efectos de interferencia de los estímulos irrelevantes permanecieron constantes, aun después de una práctica prolongada. Los hallazgos de Tipper y sus colegas sugieren que existen otros mecanismos de atención selectiva y que operan con diferentes niveles de experiencia.

Diseños multigrupales con grupos correlacionados

Varianza de las unidades

Mientras que es difícil aparear tres o cuatro conjuntos de participantes, y mientras que en la investigación del comportamiento por lo común no es factible ni deseable utilizar a los mismos participantes en cada uno de los grupos, hay situaciones naturales donde existen grupos correlacionados. Tales situaciones son particularmente importantes en la investigación educativa. Hasta hace poco, las varianzas debidas a las diferencias entre clases, escuelas, sistemas escolares y otras unidades "naturales" no se habían controlado bien o no habían sido utilizadas con la frecuencia deseada en el análisis de datos. Quizá la primera indicación de la importancia de este tipo de varianza fue dada en el magnífico libro de Lindquist (1940) sobre el análisis estadístico en la investigación educativa. En esta obra, Lindquist da un énfasis considerable a la varianza de las escuelas. Las escuelas, clases y otras unidades educativas tienden a diferir significativamente respecto al aprovechamiento, la inteligencia, las aptitudes y otras variables. El investigador educativo debe permanecer alerta ante estas ***diferencias de las unidades***, así como a las ***diferencias individuales***.

Considere un ejemplo obvio. Suponga que un investigador elige una muestra de cinco escuelas por su variedad y homogeneidad. La meta, por supuesto, es la validez externa: la representatividad. El investigador utiliza alumnos de las cinco escuelas y combina las medidas de las cinco para probar las diferencias entre medias en alguna variable dependiente. Al hacerlo, el investigador ignora la varianza debida a las diferencias entre las escuelas. Es entendible que las medias no difieran significativamente; la varianza de las escuelas está mezclada con la varianza del error.

Pueden surgir grandes errores por ignorar la varianza de las unidades tales como escuelas y clases. Uno de estos errores consiste en seleccionar varias escuelas y designar algunas de ellas como unidades experimentales y otras como unidades control. Aquí la varianza entre escuelas se enreda con la varianza de la variable experimental. De forma

similar, las clases, los distritos escolares y otras unidades educativas difieren y, por lo tanto, producen varianza. Las varianzas deben identificarse y controlarse, ya sea por medio de control experimental o estadístico, o de ambos.

Diseño factorial con grupos correlacionados

Los modelos factoriales pueden combinarse con la noción de unidades para producir un diseño valioso: el diseño *factorial de grupos correlacionados*, el cual es apropiado cuando las unidades son parte natural de la situación de investigación. Por ejemplo, la investigación quizá requiera la comparación de una variable antes y después de una intervención experimental, o antes y después de un evento importante. En efecto, habrá correlación entre las medidas antes-después de la variable dependiente. Otro ejemplo útil se presenta en la figura 21.5. Éste es un diseño factorial de 3×2 con cinco unidades (clases, escuelas, etcétera) en cada nivel de B_1 y B_2.

Las fortalezas y debilidades del diseño factorial con grupos correlacionados son similares a las de diseños factoriales más complejos. Las principales fortalezas son la habilidad para aislar y medir las varianzas y probar las interacciones. Observe que las dos principales fuentes de varianza, *tratamiento* (A) y *niveles* (B), así como las unidades de varianza, pueden evaluarse, es decir, es posible probar la significancia de las diferencias entre las medias de A, B y de las unidades. Además, pueden probarse tres interacciones: tratamientos por niveles, tratamientos por unidades y niveles por unidades. Si se utilizan puntuaciones individuales en las casillas en lugar de medias, entonces también puede probarse la interacción triple. Note lo importante que resulta dicha interacción, tanto teórica como prácticamente. Por ejemplo, se responderían preguntas como las siguientes: ¿los tratamientos operan de forma diferente en unidades distintas? ¿Ciertos métodos funcionan de manera distinta en diferentes niveles de inteligencia? ¿Con diferentes sexos? ¿Con niños de distintos niveles socioeconómicos? El estudiante avanzado deseará saber cómo manejar unidades (escuelas, clases, etcétera) y unidades de varianza en diseños factoriales. Una guía detallada se encuentra en Edwards (1984) y en Kirk (1995). El tema es difícil, incluso los nombres de los diseños se vuelven complejos: bloques aleatorizados, tratamientos anidados, diseños

FIGURA 21.5

			Métodos (tratamiento)		
		Unidades	A_1	A_2	A_3
Niveles (dispositivos, tipos, etcétera)	B_1	1	Medias de Y o medidas		
		2			
		3			
		4			
		5			
	B_2	1			
		2			
		3			
		4			
		5			

de diagrama dividido. Sin embargo, tales diseños son poderosos: combinan las virtudes de los diseños factoriales y de los diseños con grupos correlacionados. Cuando se requiera, Edwards y Kirk serán una buena guía. Además, se sugiere solicitar ayuda de alguien que entienda tanto de estadística como de investigación del comportamiento. Es absurdo utilizar programas computacionales sólo porque sus nombres parezcan apropiados o porque estén disponibles. También lo es buscar ayuda analítica del personal de informática; no es posible esperar que ellos conozcan y entiendan, por ejemplo, del análisis factorial de varianza, ya que ése no es su campo. Se tratará más sobre análisis computacional en capítulos posteriores.

Suedfeld y Rank: líderes revolucionarios y complejidad conceptual

Suedfeld y Rank (1976) probaron la intrigante noción de que los líderes revolucionarios existosos —Lenin, Cromwell y Jefferson, entre otros— son conceptualmente simples en sus discursos públicos *antes* de la revolución y conceptualmente complejos *después* de la misma. Los líderes revolucionarios no exitosos, por el otro lado, no difieren en su complejidad conceptual antes y después de la revolución. El problema se presta para un diseño factorial y para un análisis de medidas repetidas. El diseño y los datos sobre la complejidad conceptual se muestran en la tabla 21.5. Puede verse que los líderes exitosos se tornaron conceptualmente más complejos —de 1.67 a 3.65— pero los líderes no exitosos no cambiaron mucho —de 2.37 a 2.21—. La razón F de la interacción fue 12.37, significativa al nivel .005. La hipótesis fue apoyada.

Deben aclararse algunos puntos aquí. Primero, observe la combinación efectiva del diseño factorial y de las medidas repetidas. Cuando la combinación es apropiada, como en este caso, es bastante efectiva principalmente porque deja de lado, por así decirlo, la varianza en las medidas de la variable dependiente debidas a las diferencias individuales (o de grupo o de bloque). Por lo tanto, el término del error es menor y más capaz de evaluar la significancia estadística de las diferencias entre las medias. En segundo lugar, dicho estudio fue no experimental: no se manipuló ninguna variable experimental. En tercer lugar, y lo más importante, el interés intrínseco y la significancia del problema de investigación y su teoría; y la ingenuidad de medir y utilizar la complejidad conceptual como variable para "explicar" el éxito de los líderes revolucionarios ensombrece posibles puntos metodológicamente cuestionables. La frase anterior, por ejemplo, quizá sea incongruente con el uso de las variables en este estudio. Suedfeld y Rank analizaron medidas de la variable independiente, complejidad conceptual; pero la hipótesis bajo estudio en realidad era: si hay complejidad conceptual (después de la revolución), entonces habrá liderazgo exitoso. Sin embargo, con un problema de investigación de tan imponente interés y con una variable de tal importancia (complejidad conceptual) medida con gran imaginación y competencia, ¿quién quiere objetar?

TABLA 21.5 *Diseño factorial con medidas repetidas: líderes revolucionarios (estudio de Suedfeld y Rank)*[a]

	Antes de tomar el poder	Después de tomar el poder	
Éxito	1.67	3.65	2.66
Fracaso	2.37	2.22	2.30
	1.96	3.05	

[a] Las medidas en la tabla son medias de complejidad conceptual. F de la interacción = 12.37 ($p < .005$).

Perrine, Lisle y Tucker: ofrecimiento de ayuda y disposición para buscar apoyo

Los maestros en todos los niveles de educación utilizan un programa sobre la asignatura para introducir a los estudiantes al curso. ¿Qué y cuántas características del programa tienen el mayor impacto en los estudiantes, incluso antes de que empiece la instrucción en el salón de clases? Perrine, Lisle y Tucker (1995) realizaron un estudio para saber si el ofrecimiento de ayuda en el programa del instructor anima a los estudiantes universitarios, de diferentes edades, a buscar ayuda de sus instructores. De acuerdo con Perrine y sus colaboradores, éste es el primer estudio que explora el uso del apoyo social, por parte de instructores universitarios, en beneficio de los estudiantes. Perrine y sus colaboradores también estudiaron el efecto del tamaño de la clase en la disposición de los estudiantes para buscar ayuda. El estudio utilizó 104 estudiantes de licenciatura, de los cuales 82 eran mujeres y 22 eran hombres. Se pidió a cada participante leer una descripción de dos clases de psicología; las descripciones incluían afirmaciones realizadas por los instructores de cada clase en los programas de la asignatura. En la descripción se manipuló el tamaño de la clase, incluyendo 15, 45 o 150 estudiantes. El curso era descrito como demandante de mucho trabajo, pero digno de disfrutarse. También animaba a los estudiantes a no retrasarse en las lecturas ni en las tareas. Las dos afirmaciones separadas de los instructores consistieron en una que demostraba apoyo y otra que permanecía neutral. En la declaración de apoyo se animaba al estudiante a acercarse al instructor para pedir ayuda si alguna vez encontraba problemas en la clase; el neutral no incluía dicho comentario. Cada participante leyó ambas descripciones, después de lo cual, el participante respondía preguntas acerca de su disposición a buscar ayuda del instructor por seis posibles problemas académicos encontrados en la clase: 1) dificultades para entender un libro de texto, 2) baja calificación en el primer examen, 3) problemas al escuchar la exposición del instructor, 4) habilidades de estudio inefectivas para el curso, 5) planes para abandonar el curso y 6) dificultades para entender un tema importante. El participante utilizaba una escala de evaluación de 6 puntos que iba desde 0 = definitivamente no, hasta 6 = definitivamente sí.

El diseño fue un diseño factorial de 3 × 2 × 2 (tamaño de la clase × afirmaciones del discurso × edad del estudiante). El diseño contenía una variable independiente manipulada (activa), una variable independiente medida (atributo) y una variable independiente dentro de sujetos (correlacionada). El tamaño de la clase era la variable independiente manipulada y aleatorizada. La edad del estudiante era la variable independiente medida y el comentario del programa fue la variable independiente correlacionada. El uso del análisis de varianza apropiado (generalmente conocido como ANOVA mezclado, cuando al menos una varia-

TABLA 21.6 *Medias y valores F de las diferencias de los comentarios del programa y diferencias de edad (estudio de Perrine, Lisle y Tucker)*

	Programa			*Edad*		
Problema académico	**Con apoyo**	**Neutral**	***F***	**Mayores**	**Menores**	***F***
Dificultades para entender un libro de texto	4.7	3.7	76.08**	4.8	4.1	5.48*
Baja calificación en el primer examen	4.8	4.0	49.89**	5.2	4.3	7.64*
Problemas al escuchar la exposición del instructor	4.4	3.8	36.05**	4.4	4.0	1.01
Habilidades de estudio inefectivas para el curso	4.7	3.6	79.57**	4.8	4.0	6.32*
Planes para abandonar el curso	4.9	3.8	61.80**	4.8	4.3	2.18
Dificultades para entender un tema importante	5.3	4.2	82.97**	5.3	4.6	7.69*

* $p < .05$
** $p < .01$

ble independiente es entre sujetos, y al menos otra es dentro de sujetos) reveló que los participantes expresaron significativamente mayor disposición para buscar ayuda del instructor cuando la declaración de apoyo aparecía en el programa de la asignatura, que cuando sólo aparecía el comentario neutral. Los estudiantes más jóvenes (menores de 25 años) expresaron menor disposición que los estudiantes mayores. También hubo una interacción de *edad* × *programa* ($F = 4.85$, $p < .05$) que fue significativa. La respuesta al ofrecimiento de ayuda fue diferente entre los grupos de edades. Los comentarios afectaron menos a los estudiantes más jóvenes que a los mayores. El tamaño de la clase no pareció ser un factor significativo respecto a si los estudiantes estaban dispuestos o no a solicitar ayuda. La tabla 21.6 presenta el resumen estadístico del estudio.

Análisis de covarianza

La invención de Ronald Fisher del análisis de covarianza fue un evento importante en la metodología de la investigación del comportamiento. Constituye un uso creativo de los principios de la varianza, comunes al diseño experimental y a la teoría de la correlación y la regresión —que se estudiarán más adelante en el libro— para ayudar a resolver un antiguo problema del control.

El *análisis de covarianza* es una forma de análisis de varianza que prueba la significancia de las diferencias entre las medias de los grupos experimentales, después de tomar en cuenta las diferencias iniciales entre los grupos y la correlación de las medidas iniciales y las medidas de la variable dependiente. Es decir, el análisis de covarianza analiza las diferencias entre los grupos experimentales sobre Y, la variable dependiente, después de tomar en cuenta ya sean las diferencias iniciales entre los grupos sobre Y (pretest) o las diferencias entre los grupos en alguna(s) variable(s) independiente(s) potencial(es), X, correlacionadas sustancialmente con Y, la variable dependiente. La medida utilizada como variable control —el pretest o variable pertinente— se llama un *covariable*.

El lector debe ser precavido al utilizar el análisis de covarianza; es particularmente sensible a las violaciones de sus supuestos. El mal uso potencial de este método fue de tanta preocupación que la revista *Biometrics*, en 1957, dedicó un ejemplar completo a ello. Elashoff (1969) escribió un artículo importante para los investigadores educativos respecto al uso de este método. El consenso es que generalmente no es buena idea utilizarlo para diseños de investigación no experimentales.

Clark y Walberg: reforzamiento masivo y rendimiento en lectura

No tiene mucho caso describir los procedimientos y cálculos estadísticos del análisis de covarianza. Primero, porque en su forma convencional son complejos y difíciles de seguir; segundo, aquí sólo se desea mostrar el significado y propósito del método; tercero y más importante, existe una forma más fácil de hacer lo mismo que el análisis de covarianza hace. Más adelante en el libro se verá que el análisis de covarianza es un caso especial de regresión múltiple y es mucho más fácil realizarlo con las técnicas de la regresión múltiple. Para dar al lector una idea de lo que se logra con el análisis de covarianza, se estudiará un efectivo del procedimiento en estudios educativos y psicológicos.

Clark y Walberg (1968) pensaron que sus participantes, quienes posiblemente abandonarían la escuela pues su rendimiento era deficiente, necesitaban mucho más reforzamiento (ánimo, recompensa, etcétera) que los participantes que se desempeñaban bien. Por lo tanto, utilizaron reforzamiento masivo con su grupo experimental de participantes y reforzamiento moderado con su grupo control de participantes. Puesto que su variable dependiente, *rendimiento en lectura*, está altamente correlacionada con la *inteligencia*, tam-

TABLA 21.7 *Paradigma del análisis de covarianza (estudio de Clark y Walberg)*

	Experimental (reforzamiento masivo)		Control (reforzamiento moderado)	
	X (Inteligencia)	*Y* (Lectura)	*X* (Inteligencia)	*Y* (Lectura)
Medias	92.05	31.62	90.73	26.86

bién necesitaban controlar la inteligencia. Un análisis de varianza de un factor de las medias del rendimiento en lectura, de los grupos experimental y control, produjo una F de 9.52, significativa al nivel .01, lo cual apoyó su creencia. No obstante, es posible que la diferencia entre los grupos experimental y control se debiera a la inteligencia más que al reforzamiento. Es decir, aunque los sujetos fueron asignados aleatoriamente al grupo experimental, una diferencia inicial en la inteligencia, a favor del grupo experimental, pudo haber sido suficiente para volver la media de lectura del grupo experimental significativamente mayor que la media de lectura del grupo control, ya que la inteligencia está altamente correlacionada con la lectura. Con la asignación aleatoria es poco probable que suceda, pero puede ocurrir. Para controlar esta posibilidad, Clark y Walberg utilizaron el análisis de covarianza.

Estudie la tabla 21.7 que presenta un bosquejo del diseño y del análisis. Las medias de las puntuaciones de X y de Y, como reportaron Clark y Walberg, aparecen al final de la tabla. La medidas de Y son la parte más importante; resultaron significativamente diferentes. Aunque es dudoso que el análisis de covarianza cambie estos resultados, es posible que la diferencia entre las medias de X, 92.05 y 90.73 haya inclinado las balanzas estadísticas, en la prueba de la diferencia entre las medias de Y, a favor del grupo experimental. La prueba F del análisis de covarianza, que utiliza las sumas de cuadrados y los cuadrados medios de Y libres de la influencia de X, fue significativa al nivel .01: $F = 7.90$. Así, las puntuaciones promedio de lectura de los grupos experimental y control difirieron significativamente, después de ajustarlas y controlarlas respecto a la inteligencia.

Diseño y análisis de investigación: observaciones concluyentes

Cuatro objetivos principales guiaron la organización y la preparación de la parte seis de este libro. El primero fue familiarizar al estudiante respecto a los principales diseños de investigación. Al hacerlo se esperaba que se ampliaran los conceptos estrechamente circunscritos sobre la realización de investigación con, digamos, sólo un grupo experimental y sólo un grupo control; o con participantes apareados o con un grupo, antes y después. El segundo objetivo fue brindar un sentido de la estructura equilibrada de los buenos diseños de investigación, para desarrollar una idea sensible por la arquitectura del diseño. El diseño debe ser formal, así como funcional (con la y ajustada a los problemas de investigación que se busca resolver). El tercer objetivo consistió en ayudar al lector a entender la lógica de la investigación experimental y los diferentes diseños. Los diseños de investigación son rutas alternativas hacia el mismo destino: planteamientos válidos y confiables de las relaciones entre variables. Algunos diseños, si pueden llevarse a la práctica, generan planteamientos relacionales más fuertes que otros diseños.

En cierto sentido, el cuarto objetivo de la parte seis —ayudar al estudiante a comprender la relación entre el diseño de investigación y la estadística— es el más difícil de lograr.

La estadística es, en un sentido, la disciplina técnica del manejo de la varianza; y, como se ha visto, uno de los propósitos básicos del diseño consiste en proporcionar control de las varianzas sistemática y del error. Ésta es la lógica utilizada para tratar la estadística en tanto detalle en la parte cuatro y en la parte cinco, antes de considerar el diseño en la parte seis. Fisher (1951, p. 3) expresa dicha idea de forma sucinta cuando dice: "El procedimiento estadístico y el diseño experimental son sólo dos aspectos diferentes de un todo, y ese todo comprende todos los requerimientos lógicos del proceso completo de adición al conocimiento natural por medio de la experimentación."

Un diseño bien concebido no es garantía de la validez de los hallazgos de investigación. Los diseños elegantes y bien adaptados a los problemas de investigación aun pueden resultar en conclusiones erróneas o distorsionadas. Sin embargo, las oportunidades de llegar a conclusiones precisas y válidas son mayores con diseños sólidos que con aquellos que no lo son. Esto es relativamente seguro: si un diseño es inadecuado, no es factible llegar a conclusiones claras. Si, por ejemplo, se utiliza un diseño de dos grupos con sujetos apareados cuando el problema de investigación demanda lógicamente un diseño factorial, o si se utiliza un diseño factorial cuando la naturaleza de la situación de investigación requiere de un diseño de grupos correlacionados, ninguna cantidad de manipulación interpretativa o estadística puede incrementar la confianza en las conclusiones de dicha investigación.

Fisher (1951) dijo la última palabra a este respecto. En el primer capítulo de su libro, *The Design of Experiments*, afirma:

> Si el diseño de un experimento resulta inadecuado, cualquier método de interpretación que lo convierta en decisivo es inadecuado también. Es verdad que existe una enorme cantidad de procedimientos experimentales que son bien diseñados, que pueden conducir a conclusiones decisivas; pero en otras ocasiones pueden fallar en hacerlo; en tales casos, si de hecho se sacan conclusiones decisivas cuando no estén justificadas, podemos afirmar que la falla está por entero en la interpretación, no en el diseño. Pero la falla de la interpretación... reside en pasar por alto los rasgos característicos del diseño, lo que conduce a que el resultado algunas veces no sea concluyente, o concluyente en algunos aspectos pero no en todos. Comprender correctamente un aspecto del problema es entender el otro (p. 3).

Anexo computacional

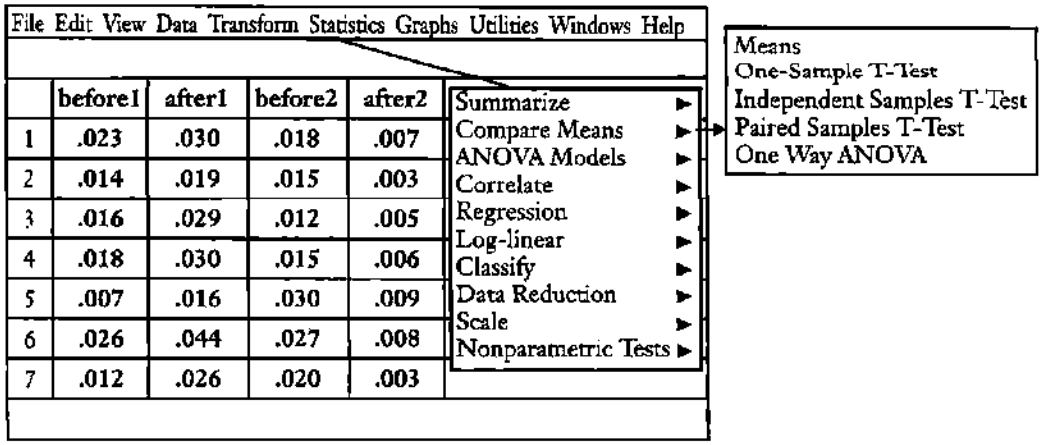
Los diseños aleatorizados pueden analizarse con pruebas *t* de muestras independientes o análisis de varianza. La organización y análisis del SPSS se incluyeron en los capítulos 13 y 14. Aquí se explicará cómo utilizar el SPSS para llevar a cabo análisis estadísticos de datos de diseños de grupos correlacionados (medidas repetidas). Para el análisis se utilizarán los datos de Miller y DiCara (1968) presentados en la tabla 21.4.

Siguiendo las instrucciones previamente establecidas sobre el ingreso de los datos, usted ingresará los datos en el SPSS, de tal manera que la hoja de datos del SPSS resultante se vea como la que aparece en la figura 21.6.

Recuerde la discusión previa, la meta era comparar ratas que presentaban un aumento en la secreción de orina con ratas cuya secreción de orina disminuía. El volumen de la secreción de orina de las ratas que mostraron un incremento es "before1" y "after1". Las variables "before2" y "after2" sirven para representar las secreciones antes y después de las ratas que mostraron una disminución.

Para que el SPSS realice el análisis apropiado para los datos presentados en la tabla 21.4 y en la figura 21.6, señale y haga clic en la opción "Statistics". Esta acción presentará un menú del cual usted debe escoger "Compare Means". Después de hacer clic en "Com-

FIGURA 21.6 *Datos de Miller y DiCara en el SPSS*



pare Means", se despliega otro menú, del cual debe elegir "Paired Sampled T- Test". Después de esta elección, aparece una nueva pantalla (véase figura 21.7).

Aquí pueden probarse simultáneamente dos pruebas *t* de muestras dependientes, siguiendo los siguientes pasos:

1. Resalte la variable "after1" (señale y haga clic en ella).
2. Resalte la variable "before1".
3. Haga clic en el botón de la flecha.
4. Resalte la variable "after2".
5. Resalte la variable "antes2".
6. Haga clic en el botón de la flecha.

FIGURA 21.7 *Pantalla del SPSS utilizada para especificar las variables para el análisis*

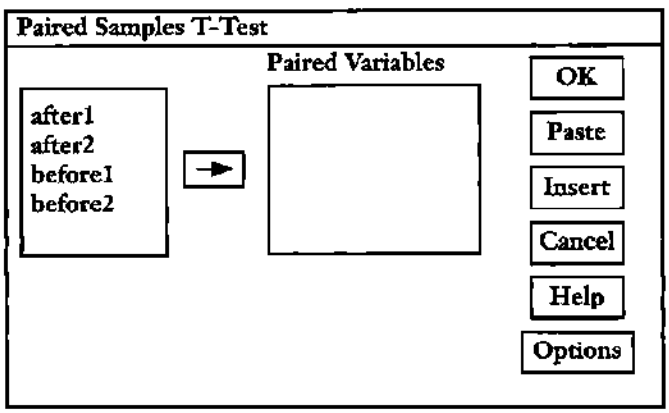
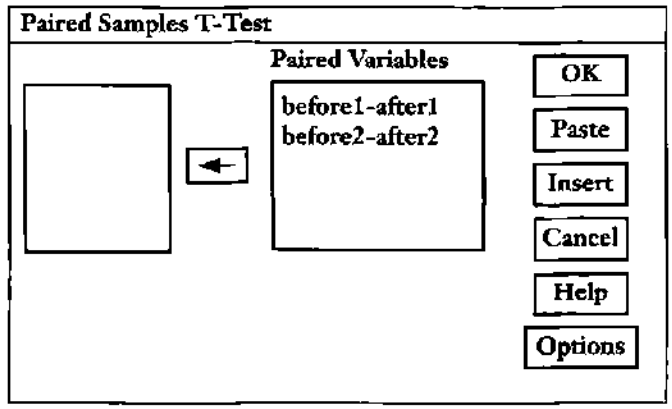


FIGURA 21.8 *Preparación para el análisis del SPSS*



En este punto usted verá que el SPSS ha formado dos ecuaciones de diferencia, que aparecen en el lado derecho del cuadro. Esto se observa en la figura 21.8. Cuando haga clic en el botón "OK", el SPSS realizará el análisis y mostrará la tabla de resultados. Una versión abreviada de la tabla de resultados se presenta en la figura 2.9.

Este análisis se realizó utilizando la prueba *t* del SPSS. Usted también puede realizar el mismo análisis con el comando "General Linear Model" del SPSS.

RESUMEN DEL CAPÍTULO

1. Los diseños de sujetos aleatorizados son los diseños preferentes para la investigación del comportamiento.
2. Los diseños de sujetos aleatorizados son verdaderos experimentos con variables independientes activas, manipuladas.
3. El método estadístico generalmente utilizado para analizar datos de diseños de sujetos aleatorizados es el análisis de varianza.
4. Los diseños de sujetos aleatorizados por lo común requieren de un número (*N*) grande de participantes para lograr la precisión deseada.

FIGURA 21.9 *Resultados del SPSS*

		Paired Differences Mean	Std. Deviation	t	df	Sig. (2-tailed)
Pair 1 Increase	BEFORE1-AFTER1	−.00111	.00445	−6.624	6	.001
Pair 2 Decrease	BEFORE2-AFTER2	.00137	.00531	6.828	6	.000

5. Los diseños de sujetos correlacionados generalmente incluyen
   a) el uso de los mismos participantes en cada condición de tratamiento
   b) aparear a los participantes en una o más variables independientes relacionadas con la variable dependiente.
   c) El empleo de más de un grupo de participantes (por ejemplo, salones de clase).
6. Las unidades pueden ser diferentes tipos de entidades. En la investigación psicológica, las unidades por lo general son personas o animales.
7. Los diseños de sujetos correlacionados incluyen al diseño de un grupo con ensayos (medidas) repetidos.
8. El diseño 20.2 es el diseño a usar cuando los participantes son apareados y fueron asignados aleatoriamente a los grupos de tratamiento.
9. Una covariable es una variable independiente potencial utilizada para ajustar las diferencias individuales entre los grupos, que *no* se deben al tratamiento. Los pretest son las covariables más comunes.
10. El análisis de covarianza es un método de sujetos correlacionados del análisis estadístico. Una covariable ajusta la variable dependiente y, después, los valores ajustados se utilizan en un análisis de varianza. La regresión múltiple es otro método estadístico que sirve para tal propósito.

SUGERENCIAS DE ESTUDIO

1. Al estudiar diseño de investigación resulta útil realizar análisis de varianza, tantos como sea posible: análisis simples de un factor y análisis factoriales de dos variables, quizá incluso un análisis de tres variables. Por medio de esta práctica estadística usted logrará un mejor entendimiento de los diseños. También puede asignar nombres de variables a sus "datos", en lugar de trabajar únicamente con números. A continuación se incluyen algunas sugerencias para realizar proyectos con números aleatorios.
   a) Obtenga tres grupos de números aleatorios del 0 al 9. Asigne nombres a las variables independiente y dependiente. Formule una hipótesis y tradúzcala a lenguaje de diseño estadístico. Realice un análisis de varianza de un factor. Interprete.
   b) Repita el paso 1 *a)* utilizando cinco grupos de números.
   c) Ahora, sume 2 a cada uno de los datos en uno de sus grupos y reste 2 a cada uno de los datos de otro grupo. Repita el análisis estadístico.
   d) Extraiga cuatro grupos de números aleatorios, con 10 números en cada uno. Ordénelos aleatoriamente en un diseño factorial de 2×2. Realice un análisis de varianza factorial.
   e) Produzca un sesgo en los números de las dos casillas derechas al sumar 3 a cada número. Repita el análisis. Compare los resultados con los del inciso *d)*.
   f) Produzca un sesgo en los números de los datos del inciso *d)* de la siguiente forma: sume 2 a cada uno de los números de las casillas superior izquierda e inferior derecha. Repita el análisis e interprete.
2. Regrese al capítulo 14, a las sugerencias de estudio 2 y 3. Trabaje ambos ejemplos de nuevo. ¿Son más fáciles para usted ahora?
3. Suponga que usted es el director de una escuela primaria. Algunos de los maestros de cuarto y quinto grado desean prescindir de los libros de trabajo. Al director general no le gusta la idea pero está dispuesto a permitirle a usted probar la idea de que los libros de trabajo no hacen mucha diferencia. (Uno de los maestros incluso sugi-

TABLA 21.8 *Datos hipotéticos (medias) de un experimento factorial ficticio*

	Métodos			
	A_1	A_2	A_3	
Hombre	45	45	36	42
Mujer	35	39	40	38
	40	42	38	

rió que los libros de texto pueden traer efectos nocivos tanto en los maestros como en los alumnos.) Para probar la eficacia de dichos libros, establezca dos planes y diseños de investigación: de un factor y otro factorial. Considere las variables *rendimiento*, *inteligencia* y *género*. También podría considerar la actitud de los maestros hacia los libros de trabajo como una posible variable independiente.

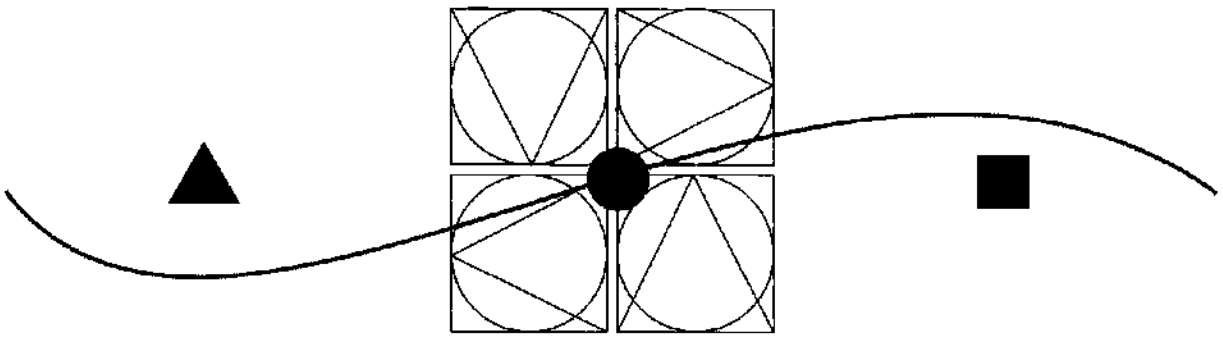
4. Suponga que se realizó una investigación que utilizó *métodos* y *género* como variables independientes y *logro* como variable dependiente, y sus resultados son los que se reportan en la tabla 21.8. Los números en las casillas son las medias ficticias. Las razones *F* de *métodos* y *género* no son significativas. La razón de *F* en la interacción es significativa al nivel .01. Interprete los resultados estadística y sustantivamente. Para hacer esto último, asigne nombres a los tres métodos.
5. Aunque difícil y en ocasiones frustrante, no existe un sustituto para la lectura y el estudio de reportes de investigación originales. En éste y en capítulos previos se han citado y resumido numerosos estudios que utilizan un diseño factorial y el análisis de varianza. Seleccione y lea dos de esos estudios e intente resumir alguno. Critique ambos estudios respecto a la adecuación del diseño y la realización de la investigación (con lo mejor de su conocimiento y habilidades actuales). Enfóquese particularmente en la adecuación del diseño para responder la(s) pregunta(s) de investigación.
6. Se realizó un análisis de varianza de dos factores (medidas repetidas) con los datos de Miller y DiCara sobre las ratas con incremento, en la tabla 21.4, con algunos de los datos reportados en la tabla: ω^2 (omega al cuadrado de Hays) fue de .357; ω^2 para los datos de las ratas con disminución fue de .633. ¿Qué significan estos coeficientes? ¿Por qué calcularlos?
7. Kolb (1965), quien basó su estudio en el trabajo sobresaliente de McClelland sobre motivación de logro, realizó un experimento fascinante con jóvenes de secundaria con bajo logro y alta inteligencia. De 57 jóvenes, asignó a 20 aleatoriamente a un programa de entrenamiento donde, a través de distintos medios, se les "enseñó" motivación de logro a los jóvenes (un intento por crear una necesidad de logro en los jóvenes). A los jóvenes se les aplicó un pretest de motivación de logro en el verano, el cual se aplicó otra vez seis meses después. Las *puntuaciones* medias de cambio fueron, para los grupos experimental y control, 6.72 y –.34, respectivamente. Éstas fueron significativas al nivel .005.
   a) Comente sobre el uso de puntuaciones de cambio. ¿Su uso debilita la fe que usted tiene en la significancia estadística de los resultados?
   b) ¿Pueden otros factores, diferentes del entrenamiento experimental, haber inducido el cambio? Si así es, ¿cuáles serían esos factores?
8. Para evitar que el estudiante crea que sólo se analizan medidas continuas y que el análisis de varianza sólo se utiliza en experimentos psicológicos y educativos, lea el estudio de Freedman, Wallington y Bless (1967) sobre la culpa y la sumisión. Había

un grupo experimental (sujetos inducidos a mentir) y un grupo control. La variable dependiente se midió viendo si un participante obedecía o no a una solicitud de ayuda. Los resultados fueron reportados en tablas de frecuencias de tabulación cruzada. Lea el estudio y, después de estudiar el diseño y los resultados de los autores, diseñe uno de los tres experimentos de otra forma. Introduzca otra variable independiente, por ejemplo. Suponga que se sabía que había grandes diferencias individuales en la sumisión. ¿Cómo puede controlarse lo anterior? Asigne nombre y describa dos tipos de diseño para hacerlo.

9. En un estudio donde el entrenamiento en las complejidades de los estímulos artísticos afectó la actitud hacia la música, entre otras cuestiones, Renner (1970) utilizó un análisis de covarianza, donde la covariable eran las medidas de una escala diseñada para medir la actitud hacia la música. Éste fue el pretest. Hubo tres grupos experimentales. Estructure el diseño a partir de esta breve descripción. ¿Por qué Renner utilizó la escala de actitud hacia la música como pretest? ¿Por qué utilizó un análisis de covarianza? (Nota: vale la pena leer el reporte original. El estudio, que en parte trata sobre creatividad, es creativo en sí.)
10. En un estudio significativo del efecto de la educación en artes liberales sobre la formación de conceptos complejos, Winter y McClelland (1978) encontraron que la diferencia entre los estudiantes de primer y segundo año de una universidad de artes liberales, respecto a la medida de la formación de conceptos complejos, fue estadísticamente significativa ($M_{\text{primer año}} = 2.00$, $M_{\text{segundo año}} = 1.22$; $t = 3.76$; ($p < .001$). Como se dieron cuenta de que se necesitaba una comparación, también probaron las diferencias de medias similares en una universidad pedagógica y una universidad comunitaria. Ninguna de estas diferencias resultó estadísticamente significativa. ¿Por qué Winter y McClelland probaron la relación en la universidad pedagógica y en la universidad comunitaria? Se sugiere que los estudiantes encuentren y lean el reporte original —vale la pena su estudio— y realicen un análisis de varianza de las n, medias y desviaciones estándar reportadas, utilizando el método descrito en el capítulo 13 (anexo).
11. Una virtud del análisis de covarianza, rara vez mencionada en los textos, es que pueden calcularse tres estimados de la correlación entre X y Y. Éstos son: (i) la r total sobre todas las puntuaciones; (ii) la r entre grupos, que es la r entre las medias de X y de Y; (iii) y la r dentro de grupos, la r calculada a partir de un promedio de las r entre X y Y dentro de k grupos. La r dentro de grupos es el "mejor" estimado de la r "verdadera" entre X y Y. ¿Por qué es esto así?
[Sugerencia: ¿Puede una r total, aquella que se calcula generalmente en la práctica, inflarse o desinflarse por la varianza entre grupos?]

Parte Siete

Tipos de investigación

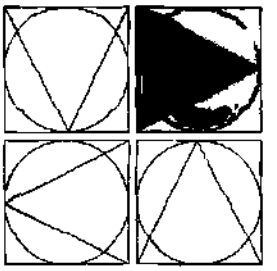


Capítulo 22
Diseños de investigación cuasi-experimentales y con $N = 1$

Capítulo 23
Investigación no experimental

Capítulo 24
Experimentos de laboratorio, experimentos de campo y estudios de campo

Capítulo 25
Investigación por encuesta



CAPÍTULO 22

DISEÑOS DE INVESTIGACIÓN CUASI-EXPERIMENTALES Y CON $N = 1$

En capítulos previos se estableció y enfatizó que una de las principales metas de la ciencia consiste en encontrar relaciones causales. En las ciencias del comportamiento, el experimento verdadero es la técnica más fuerte utilizada para alcanzar dicha meta. Cuando el experimento verdadero se estructura y ejecuta correctamente, proporciona al investigador una proposición de *causa y efecto* respecto a la relación entre X (variable independiente) y Y (variable dependiente). Esta se considera generalmente la forma más elevada de experimentación. No obstante, existen problemas de investigación en las ciencias del comportamiento, y especialmente en la investigación educativa, que no pueden estudiarse utilizando un diseño experimental verdadero; es decir, los diseños 20.1 a 20.6 y algunas de sus variantes, revisadas en los capítulos 20 y 21, no pueden utilizarse. Faltan uno o más

de los componentes de un experimento verdadero o se han debilitado, ya sea por la naturaleza del estudio o por una planeación pobre. El debilitamiento de los componentes del experimento verdadero constituye el tema que se analizará en el presente capítulo. Se examinarán dos tipos de diseños de investigación donde se ven comprometidos uno o más de los componentes del experimento verdadero. El primer tipo son los llamados diseños *cuasi-experimentales* y el segundo tipo son los conocidos como diseños *de un solo sujeto* o $N = 1$.

Diseños comprometidos, también conocidos como diseños cuasi-experimentales

Es posible, y de hecho necesario, utilizar diseños que estén comprometidos con la experimentación verdadera. Recuerde que la experimentación verdadera requiere por lo menos de dos grupos, uno que reciba un tratamiento experimental, y otro que no lo reciba o que lo reciba de forma diferente. El experimento verdadero requiere la manipulación de por lo menos una variable independiente, la asignación aleatoria de los participantes a los grupos y la asignación aleatoria del tratamiento a los grupos. Cuando falta uno o más de estos prerrequisitos por cualquier razón, se tiene un *diseño comprometido*. Los diseños comprometidos se conocen popularmente como diseños cuasi-experimentales. Se les llama *cuasi* porque dicho término significa "casi" o "tipo de". Cook y Campbell (1979) presentan dos principales clasificaciones del diseño cuasi-experimental. El primero se llama "diseño de grupo control no equivalente"; el segundo es el "diseño de series interrumpidas". Numerosos estudios de investigación que se realizan fuera del laboratorio podrían caer en una de tales categorías. Muchos estudios de investigación de mercado tienen la forma de diseños cuasi-experimentales. Con frecuencia se le pide a un investigador que "diseñe" y analice los datos de un estudio sin planeación. Por ejemplo, un comprador de abarrotes decide abastecer una nueva marca de alimento para bebés. Sus superiores se preguntan posteriormente si tal movimiento fue rentable; entonces el comprador consulta a un investigador de mercado para determinar lo que puede hacerse para demostrar si la decisión fue rentable o no. Dicho análisis no tendría la mejor selección ni asignación aleatorias, sólo consistiría de datos tomados a través del tiempo. Además, otros anuncios o la estación del año podrían influir en las ventas del alimento para bebés. El único componente que asemeja un experimento verdadero es el hecho de que se manipuló la variable independiente. No todas las tiendas recibieron ese producto en particular. Con tales problemas, el investigador optaría por el uso de diseños de investigación cuasi-experimentales o comprometidos.

Diseño de grupo control no equivalente

Quizás el diseño cuasi-experimental más utilizado es el de grupo experimental-grupo control, en el cual no se tiene mucha seguridad de que los grupos experimental y control sean equivalentes. Algunos autores como Cook y Campbell (1979), Christensen (1977), Ray (1977), y Graziano y Raulin (1993) se refieren a él como diseño de grupo control no equivalente. Cook y Campbell presentan ocho variaciones de este diseño, que ellos consideran "interpretables":

- diseños de grupo control sin tratamiento
- diseños de variables dependientes no equivalentes

diseños de grupo con retiro del tratamiento
diseños de tratamiento repetido
diseños de grupo control no equivalente con reversión del tratamiento
diseños cohorte
diseños sólo con postest
diseños de continuidad de regresión

En este libro se analizará en detalle sólo uno de ellos. Es el que tiene mayor posibilidad de ocurrir en alguna forma y variación en la literatura de investigación. Para un estudio más detallado de estos ocho tipos de diseños de grupo control no equivalente, se recomienda leer a Cook y Campbell (1979).

Diseño de grupo control sin tratamiento

La estructura del diseño de grupo control sin tratamiento ya se consideró en el diseño 20.3. Cook y Campbell (1979) se refieren a éste como el diseño de grupo control sin tratamiento con pretest y postest. La forma comprometida es como sigue:

Diseño 22.1: Diseño de grupo control sin tratamiento

Y_a	X	Y_d	(Experimental)
Y_a	$\sim X$	Y_d	(Control)

La diferencia entre el diseño 20.3 y el diseño 22.1 es marcada. En el diseño 22.1 no hay una asignación aleatorizada de los participantes a los grupos como en el 20.3(*a*), ni hay apareamiento de los participantes y luego asignación aleatoria como en el 20.3(*b*). Por lo tanto, el diseño 22.1 está sujeto a las debilidades debidas a la posible falta de equivalencia entre los grupos en variables distintas a X. Por lo común los investigadores sufren para establecer equivalencia por otros medios y, dependiendo del grado en que sean exitosos al hacerlo, el diseño será válido, lo cual se logra en formas que se analizarán a continuación.

En ocasiones es difícil o imposible igualar grupos por medio de selección o asignación aleatorias, o por medio del apareamiento. ¿Debe entonces renunciarse a llevar a cabo la investigación? Por ningún motivo. Deben realizarse todos los esfuerzos posibles para 1) seleccionar y 2) asignar aleatoriamente. Si ambas cuestiones no son posibles, quizá se puedan lograr el apareamiento y la asignación aleatoria. Si el apareamiento y la asignación aleatoria no son posibles, por lo menos debe hacerse el esfuerzo de utilizar muestras que provengan de la misma población o muestras que sean lo más similares posibles. Los tratamientos experimentales deben asignarse aleatoriamente y después debe verificarse la similitud de los grupos, utilizando cualquier información disponible (sexo, edad, clase social, etcétera). La equivalencia de los grupos puede verificarse utilizando las medias y las desviaciones estándar de los pretest: las pruebas-*t* y las pruebas-*F* sirven para este fin. Las distribuciones también deben verificarse. Aunque no se alcanza la seguridad ofrecida por la aleatorización, si todos estos aspectos resultan satisfactorios, entonces se puede continuar con un estudio, sabiendo por lo menos que no existe evidencia conocida en contra del supuesto de equivalencia.

Estas precauciones incrementan las posibilidades de conseguir validez interna. Aún existen dificultades, todas las cuales están subordinadas a una dificultad principal: la *selección*. Estas otras dificultades no se estudiarán aquí; para un análisis más detallado, véase Campbell y Stanley (1963), o Cook y Campbell (1979).

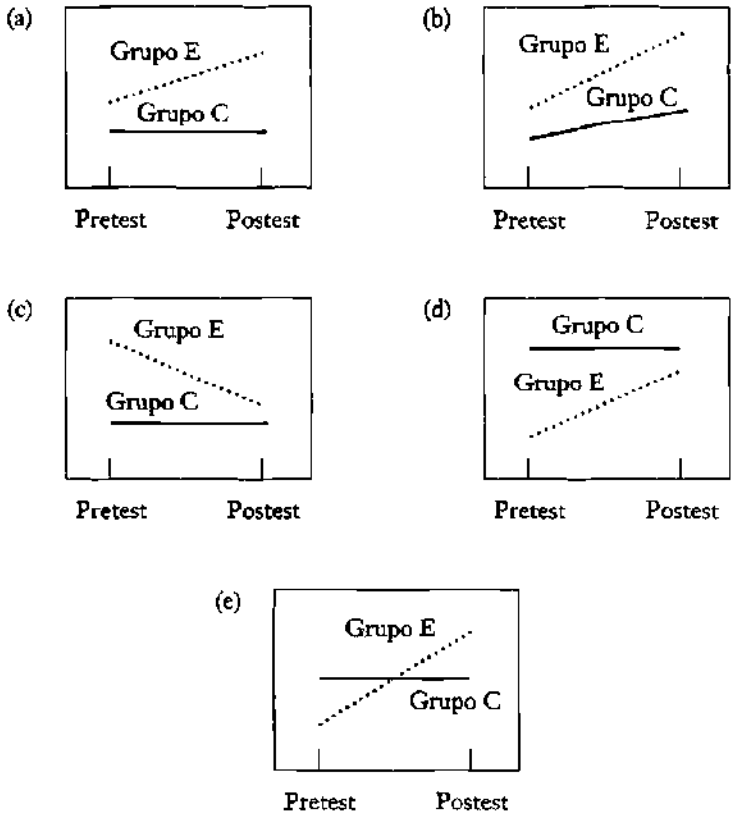
La selección constituye uno de los problemas más difíciles y complicados de la investigación del comportamiento. Puesto que sus aspectos se verán con detalle en el capítulo 23 que se trata sobre la investigación no experimental, aquí solamente se incluirá una breve descripción. Una de las razones importantes del énfasis en la selección y asignación aleatorias es evitar las dificultades de la selección. Cuando se integra a los participantes a los grupos con bases extrañas a los propósitos de investigación, a esto se le llama "selección" o, alternativamente, "autoselección". Considere un ejemplo común: suponga que los voluntarios se utilizan en el grupo experimental y otros participantes sirven como controles. Si los voluntarios difieren en una característica relacionada con *Y*, la variable dependiente, la diferencia última entre los grupos experimental y control quizá se deba a dicha característica, más que a *X*, la variable independiente (tratamiento). Los voluntarios pueden ser más (o menos) inteligentes que los no voluntarios. Si se realizara un experimento con cierto tipo de aprendizaje como variable dependiente, los voluntarios obviamente se desempeñarían mejor en *Y* debido a una inteligencia superior, a pesar de la semejanza inicial de los dos grupos en el pretest. Note que si se hubieran utilizado sólo voluntarios que se hubiesen asignado aleatoriamente a los grupos experimental y control, la dificultad de selección se disminuiría. Sin embargo, la validez externa o representatividad habría disminuido.

Cook y Campbell (1979) afirman que aun en los casos muy extremos es posible extraer conclusiones sólidas si se consideran y justifican todas las amenazas en contra de la validez. Sin el beneficio de la asignación aleatoria, deben llevarse a cabo intentos con otros medios para eliminar hipótesis rivales. Aquí se considera únicamente el diseño que utiliza el pretest debido a que éste ofrece información útil respecto a la efectividad de la variable independiente sobre la variable dependiente. El pretest puede proporcionar datos respecto a la igualdad entre los grupos, antes de la administración del tratamiento al grupo experimental.

Otro ejemplo más frecuente en investigación educativa consiste en tomar algunos grupos escolares para el grupo experimental y otros para el grupo control. Si se selecciona un número bastante grande de grupos, y se asignan aleatoriamente a los grupos experimental y control, entonces no hay mucho problema; pero si no se asignan aleatoriamente, algunos de ellos pueden asignarse a sí mismos a los grupos experimentales, y estos grupos quizá tengan características que los predisponen a tener puntuaciones medias de *Y* más altas que los otros grupos. Por ejemplo, sus maestros pueden estar más alertas, ser más inteligentes y más agresivos. Las características interaccionan con la selección para producir, independientemente de *X*, puntuaciones más altas para el grupo experimental que para el grupo control *Y*. En otras palabras, algo que influya en el proceso de selección (por ejemplo, participantes voluntarios), también influye en las medidas de la variable dependiente. Esto sucede aunque el pretest muestre que los grupos son iguales o similares respecto a la variable dependiente. La manipulación de *X* es "efectiva" debido a la selección o autoselección; pero no es efectiva por sí misma. Además, en ocasiones un investigador educativo necesita recibir la aprobación del distrito escolar para realizar la investigación. En ocasiones el distrito asignará las escuelas y los grupos que el investigador pueda usar.

Un estudio clásico de Sanford y Hemphill (1952), reportado por Campbell y Stanley (1963), utilizó este diseño. Este estudio se condujo en la U.S. Naval Academy en Annapolis, con el fin de saber si un curso de psicología en el currículum incrementaba la confianza de los estudiantes (guardias marinos) en las situaciones sociales. Los guardias marinos de segundo año fueron el primer grupo de estudiantes en tomar el curso de psicología. El grupo comparativo o control lo conformó la clase de tercer año, quienes no habían tomado el curso durante su segundo año. Se administró un cuestionario de situaciones sociales a ambas clases al inicio del año académico y al final del año. Los resultados demostraron

FIGURA 22.1 *Cinco posibles resultados del diseño de grupo control no equivalente*[a]



[a] E = grupo experimental y C = grupo control

un incremento en las puntuaciones de confianza en la clase de segundo año, de 43.26 a 51.42. La clase de tercer año también mostró un incremento; pero éste fue considerablemente menor, con un cambio de 55.80 a 56.78. A partir de estos datos se podría concluir que tomar el curso de psicología tuvo un efecto de incremento en la confianza de los sujetos en situaciones sociales. Sin embargo, también son posibles otras explicaciones. Una podría explicar que las mayores ganancias logradas por la clase de segundo año fueron el resultado de algún desarrollo maduracional que tiene su mayor crecimiento en el segundo año, y un menor crecimiento en el tercer año. Si dicho proceso existe, el mayor incremento en las puntuaciones de la clase de segundo año se hubiera dado aun si los guardias marinos no hubieran tomado la clase de psicología. El hecho de que la clase de segundo año iniciara con una puntuación más baja que la clase de tercer año, podría indicar que estos estudiantes no habían alcanzado todavía un nivel equivalente al de la clase de tercer año. Además, las puntuaciones del final del año de la clase de segundo año no fueron equivalentes a las puntuaciones iniciales de la clase de tercer año. Un mejor y más fuerte diseño consistiría en crear dos grupos equivalentes de la clase de segundo año, a

través de la selección aleatoria, e impartir la clase de psicología aleatoriamente a un solo grupo.

Resultados posibles de tal diseño se presentan en la figura 22.1. Existe la posibilidad de una interpretación diferente de la causalidad, según el resultado que obtenga el investigador. En la mayoría de los casos la amenaza más probable contra la validez interna sería la interacción selección-maduración. Quizá se recuerde que dicha interacción ocurre cuando 1) dos grupos son diferentes desde el inicio, de acuerdo a las medidas; 2) uno de los grupos experimenta mayores cambios diferenciales, como tornarse más experimentado, más preciso, más cansado, etcétera, que el otro grupo. La diferencia posterior al tratamiento, de acuerdo al postest, no puede atribuirse exactamente al tratamiento por sí mismo.

En la figura 22.1(a) existen tres amenazas posibles contra la validez interna. Como antes se mencionó, la amenaza con mayor prevalencia es la interacción selección-maduración. Para el resultado en la figura 22.1(a), Cook y Campbell (1979) afirman que hay cuatro explicaciones alternativas.

La primera es la interacción selección-maduración. Digamos que el estudio implica la comparación de dos estrategias o métodos de solución de problemas. El grupo E posee mayor inteligencia que el grupo C. El grupo E tiene puntuaciones mayores en el pretest que el grupo C. El grupo E muestra un incremento en las puntuaciones del postest después del tratamiento. El grupo C presenta poco o ningún cambio. Quizá parezca que el tratamiento que recibe el grupo E es superior al tratamiento recibido por el grupo C. Sin embargo, con la interacción selección-maduración, el incremento del grupo E puede deberse a su mayor nivel de inteligencia. Con un nivel más alto de inteligencia, tales participantes quizá pueden procesar más o quizá crezcan más rápido que los del grupo C.

Una segunda explicación se refiere a la instrumentación. La escala utilizada para medir la variable dependiente tal vez sea más sensible en ciertos niveles que en otros. Como ejemplo considere los percentiles, los cuales tienen una ventaja sobre las puntuaciones en bruto, pues transmiten un significado directo sin otras piezas de información. No obstante, los percentiles son transformaciones no lineales de las puntuaciones en bruto. En una distribución normal, los cambios en las puntuaciones en bruto cercanas al centro de la distribución reflejan cambios percentilares más grandes que en las colas. Un cambio de sólo 2 o 3 puntos en la escala de puntuación en bruto puede reflejar un cambio percentilar de 10 puntos cerca del centro de la distribución. Éste no sería el caso al considerar las colas de la distribución normal. Se necesitaría un cambio de 15 puntos de la puntuación en bruto para tener un incremento de 10 puntos percentilares en la cola de la distribución. Por lo tanto, las mediciones de los percentiles del grupo C pueden cambiar poco debido a que las mediciones no son lo suficientemente sensibles para detectar los cambios [en las colas]. Sin embargo, el grupo E mostrará una cantidad mayor de cambio, ya que su percentil está en la parte más sensible de la escala de medición.

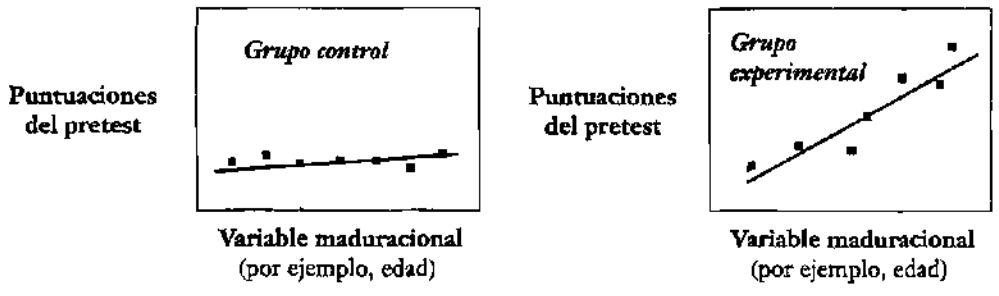
La tercera explicación es la regresión estadística. Digamos que los dos grupos, E y C, en realidad provienen de diferentes poblaciones, y que el grupo C es el grupo de interés. El investigador desea introducir un plan educativo para ayudar a incrementar el funcionamiento intelectual de estos participantes, quienes son seleccionados porque generalmente obtienen puntuaciones bajas en pruebas de inteligencia. El investigador crea un grupo comparativo o control a partir de estudiantes con puntuaciones normales. Este grupo se representa como grupo E en la figura 22.1(a). Estos estudiantes estarían en el extremo bajo de la escala de puntuaciones de la prueba; pero no tan abajo como el grupo C. Si ésta es la situación, entonces la regresión estadística constituye una explicación alternativa viable. El incremento de las puntuaciones del grupo E se deberían a su selección con base en las puntuaciones extremas. En el postest sus puntuaciones aumentarían pues estarían acercándose a la línea base de la población.

La cuarta explicación se centra en la interacción entre la historia y la selección. Cook y Campbell (1979) se refieren a lo anterior como el efecto de la historia local. En dicha situación, algo diferente a la variable independiente afectará a uno de los grupos (grupo E) y no al otro grupo (grupo C). Suponga que un investigador de mercado desea determinar la efectividad de un anuncio de condimento para sopa. Se reúnen datos sobre las ventas antes y después de la introducción del anuncio. Se utilizan dos grupos de diferentes regiones del país: un grupo es del sur de California y el otro es del oeste medio. En este caso, el crecimiento en las ventas observado en uno de los grupos (E) pudo no deberse necesariamente al anuncio. Ambos grupos pueden tener conductas de compra similares durante la primavera y el verano; es decir, que no hay una alta necesidad de los condimentos para sopa. No obstante, conforme se acerca la temporada de otoño, la venta de condimentos para sopa puede aumentar en el grupo del medio oeste. En el sur de California, donde las temperaturas son considerablemente más cálidas durante todo el año, la demanda de condimentos para sopa permanecería bastante constante. Por lo tanto, la explicación del incremento en las ventas en el medio oeste sería la estación del año y no el anuncio.

Todas las amenazas mencionadas respecto a la figura 22.1(a) también resultan verdaderas para la figura 22.1(b). Mientras que en la figura 22.1(a) uno de los grupos (grupo C) permanece constante, en la figura 22.1(b) ambos grupos experimentan un incremento del pretest al postest. La interacción selección-maduración aún es una posibilidad ya que, por definición, los grupos están creciendo (o disminuyendo) a diferente ritmo, pues el grupo de puntuaciones más bajas (grupo C) progresa a un ritmo menor que el grupo de altas puntuaciones (grupo E). Para determinar si la selección y la maduración juegan un papel importante en los resultados, Cook y Campbell (1979) recomiendan dos métodos. El primero implica observar únicamente los datos del grupo experimental (grupo E). Si la varianza dentro de grupos del postest es considerablemente mayor que la varianza dentro de grupos del postest, entonces hay evidencia de una explicación por la interacción selección-maduración. El segundo método consiste en trazar dos gráficas y la línea de regresión asociada con cada gráfica. Una gráfica es para el grupo experimental (grupo E). Las puntuaciones del pretest se grafican contra la variable de maduración, que puede ser la edad o la experiencia. La segunda gráfica sería igual, excepto que sería para el grupo control (grupo C). Si las pendientes de la línea de regresión de cada gráfica difieren entre sí, entonces existe evidencia de un ritmo promedio de crecimiento diferencial, lo cual significaría la posibilidad de una interacción selección-maduración (véase figura 22.2).

El resultado que se presenta en la figura 22.1(c) se encuentra con mayor frecuencia en los estudios de psicología clínica. Se supone que el tratamiento resultará en una disminución de una conducta indeseable. Como los dos resultados previos, éste también es susceptible de la interacción selección-maduración, de la regresión estadística, de la

FIGURA 22.2 *Comparación del grupo experimental y el grupo control*



instrumentación y de los efectos de la historia local. En este resultado, la diferencia entre los grupos experimental y control es muy dramática en el pretest, pero después del tratamiento los grupos se acercan entre sí.

El cuarto resultado se presenta en la figura 22.1(d). Éste difiere de los tres previos en que el grupo control (grupo C) inicia más alto que el grupo experimental (grupo E) y permanece más alto incluso en el postest. Sin embargo, el grupo E mostró un mayor incremento del pretest al postest. La regresión estadística sería una amenaza si los participantes del grupo E fuesen seleccionados con base en su puntuación extremadamente baja. Sin embargo, Cook y Campbell (1979) afirman que la amenaza de la selección-maduración puede excluirse, ya que este efecto generalmente resulta en un ritmo de crecimiento más lento en aquellos con puntuaciones bajas, y en un ritmo de crecimiento más rápido en aquellos con puntuaciones altas. Aquí los participantes con bajas puntuaciones muestran mayor ganancia en las puntuaciones que los participantes con puntuaciones altas. Esta evidencia confiere apoyo a la efectividad de la condición de tratamiento recibida por el grupo E. Lo que no puede eliminarse fácilmente son las amenazas de la instrumentación y de la historia local, vistas en los tres resultados previos de los diseños de grupo control no equivalente.

Con el resultado final mostrado en la figura 22.1(e), las medias de los grupos experimental (grupo E) y control (grupo C) son significativamente diferentes entre sí, tanto en el pretest como en el postest. No obstante, las diferencias resultan en dirección opuesta en el pretest, respecto al postest. Las líneas se cruzan entre sí. El grupo E inicia abajo, pero después supera al grupo C, que inicialmente tuvo puntuaciones altas. Cook y Campbell (1979) consideraron este resultado más interpretable que los cuatro anteriores. La instrumentación o la escalación se descarta, ya que ninguna transformación de las puntuaciones podría suprimir o reducir este cruce o efecto de interacción. La regresión estadística se vuelve insostenible porque es extremadamente raro que una puntuación baja tenga suficiente regresión para superar una puntuación inicialmente alta. Además de un efecto muy complicado de interacción de selección-maduración, dicho patrón no se asemeja a las amenazas por selección-maduración. La maduración, por ejemplo, generalmente no inicia diferente, se cruza y luego continúa en la dirección opuesta. Por lo tanto, el resultado de la figura 22.1(e) parece ser el más fuerte y debe permitir al investigador hacer proposiciones causales respecto al tratamiento. Sin embargo, Cook y Campbell advierten que los investigadores no deben planear el desarrollo de la investigación cuasi-experimental con la esperanza de obtener este resultado. En definitiva el diseño de un estudio de grupo control no equivalente debe realizarse con cuidado y precaución.

Ejemplos de investigación

Nelson, Hall y Walsh-Bowers: diseño de grupo control no equivalente

La investigación realizada por Nelson, Hall y Walsh-Bowers (1997) afirma específicamente que ***utilizaron un diseño de grupo control no equivalente*** para comparar los efectos a largo plazo de los apartamentos de apoyo (AA), hogares de grupo (HG) y los hogares de alojamiento y cuidado (HAC) para residentes psiquiátricos. Los apartamentos de apoyo y los hogares de grupo son manejados por organizaciones no lucrativas; los hogares de alojamiento y cuidado sí se manejan con fines lucrativos. El objetivo principal fue comparar los dos grupos de intervención: los apartamentos de apoyo y los hogares de grupo. Los autores no pudieron asignar aleatoriamente a los participantes a los diferentes hogares. Nelson *et al.*, realizaron su mayor esfuerzo para aparear a los residentes; pero existían algunas diferencias significativas en la composición de los grupos que los condujeron a

utilizar el diseño de grupo control no equivalente. Con este diseño decidieron usar a residentes de HAC como grupo comparativo. No fueron capaces de corregir, por medio del apareamiento, las siguientes variables que pudieron tener un efecto sobre las variables dependientes: 1) Los grupos AA y HG tendían a ser más jóvenes que el grupo HAC (33 años contra 45 años) y tenían menos tiempo de residencia (2.5 años contra 39 años). 2) Los residentes de AA y HG tenían un mayor nivel de educación que los del grupo de HAC. Nelson y sus colaboradores encontraron una diferencia significativa entre estos grupos en esas variables. Aunque el género no resultó significativo, había más hombres que mujeres en los grupos de AA y HG; y más mujeres que hombres en el grupo HAC.

Nelson *et al.* proponen que las diferencias que encontraron entre estos tres grupos en las medidas del postest, quizá se deban al problema de selección y no al tipo de instalación para el cuidado.

Chapman y McCauley: cuasi-experimento

En este estudio, Chapman y McCauley (1993) examinaron el desarrollo de la carrera de estudiantes graduados que solicitaron una beca para graduados de la National Science Foundation (NSF) Graduate Fellowship Award. Aunque quizás se puede pensar que este estudio es no experimental, Chapman y McCauley consideraron que podía clasificarse como cuasi-experimental; debe entenderse por qué. Al comparar a los ganadores de la beca con los no ganadores, la elección de ganadores no se realizó exactamente al azar. El estudio no consideró a los solicitantes del grupo de calidad 1. Los solicitantes del grupo 1 estaban dentro del 5% más alto y todos recibieron becas. Los solicitantes al NSF del grupo de calidad 2 conformaron el siguiente 10% y fueron considerados como un grupo altamente homogéneo. Las becas fueron otorgadas aproximadamente a la mitad del grupo homogéneo de solicitantes, por medio de un procedimiento que Chapman y McCauley consideran como una asignación aproximadamente aleatoria de la beca o de la mención honorífica. Se asignó a los estudiantes con respecto al potencial académico. Chapman y McCauley consideraron que las diferencias en el desempeño entre los solicitantes del grupo de calidad 2, donde algunos estudiantes fueron adjudicados con una beca NSF y otros no, podrían revelar el efecto de las expectativas positivas asociadas con esta prestigiosa beca.

Los resultados mostraron que quienes recibieron la beca NSF tenían mayores posibilidades de terminar el doctorado. Sin embargo, Chapman y McCauley no encontraron un efecto confiable de la beca sobre el logro de estatus dentro de la facultad o sobre el solicitar o recibir un reconocimiento del NSF o una beca de investigación de los institutos nacionales de salud. Parece que las expectativas positivas asociadas con esta prestigiosa beca tienen alguna influencia en las escuelas de posgrado; pero no la tienen en los logros posteriores a la escuela de posgrado.

Diseños de tiempo

Variantes importantes del diseño cuasi-experimental básico son los diseños de tiempo. La forma del diseño 20.6 puede alterarse para incluir un lapso de tiempo:

Y_a	X	Y_d	
Y_a	$\sim X$	Y_d	
	X		Y_d
	$\sim X$		Y_d

Las Y_4 de la tercera y cuarta líneas son observaciones de la variable dependiente en cualquier fecha posterior específica. Dicha alteración, por supuesto, cambia el propósito del diseño y puede causar que se pierdan algunas de las virtudes del diseño 20.6. Es posible, si se tiene el tiempo, la paciencia y los recursos, mantener todos los beneficios anteriores y aun extender el tiempo añadiendo dos o más grupos al propio diseño 20.6.

Un problema común de investigación, especialmente en estudios sobre el desarrollo y crecimiento de los niños, incluye el estudio de individuos y de grupos utilizando el tiempo como variable. Éstos son estudios longitudinales de los participantes, con frecuencia niños, en diferentes puntos de tiempo. Un diseño entre muchos podría ser:

Diseño 22.2: Un diseño de tiempo longitudinal (también conocido como diseño de series de tiempo interrumpidas)

Y_1	Y_2	Y_3	Y_4	X	Y_5	Y_6	Y_7	Y_8

Observe la similitud con el diseño 19.2, donde un grupo es comparado consigo mismo. El uso del diseño 22.2 permite evitar una de las dificultades del diseño 19.2; su empleo hace posible separar los efectos reactivos de medición del efecto de X. Permite determinar si las mediciones tienen un efecto reactivo y si X fue lo suficientemente fuerte para superar tal efecto. El efecto reactivo debe mostrarse a sí mismo al comparar Y_3 con Y_4; lo cual puede contrastarse con Y_5. Si existe un incremento en Y_5 por encima del incremento en Y_4, a partir de Y_3, puede atribuirse a X. Un argumento similar se aplica para la maduración y la historia.

Una dificultad con los estudios longitudinales o de tiempo, especialmente con niños, es el crecimiento o el aprendizaje que ocurre de manera natural a través del tiempo: los niños no detienen su crecimiento ni su aprendizaje para conveniencia de la investigación. A mayor periodo, mayor será el problema. En otras palabras, el tiempo en sí mismo es una variable. Con un diseño como el 20.2, Y_a X Y_b, la variable tiempo puede confundir a X, la variable independiente experimental. Si existe una diferencia significativa entre Y_a y Y_b, no es posible decir si X o una "variable" de tiempo provocó el cambio. Pero con el diseño 22.2 se tienen otras medidas de Y y, por lo tanto, una línea base con la cual comparar el cambio en Y, presumiblemente debido a X.

Un método para determinar si el tratamiento experimental tuvo un efecto consiste en observar una gráfica de los datos a través del tiempo. Caporaso (1973) presenta varios patrones de conducta adicionales posibles, los cuales se obtienen de datos de series de tiempo. Ya sea que un cambio significativo en la conducta venga o no después de la introducción de la condición de tratamiento, esto se determina por medio de una prueba de significancia. La prueba estadística más utilizada es ARIMA (promedio autorregresivo, integrado y móvil), desarrollada por Box y Jenkins (1970) (véase también Gottman, 1981). Este método consiste en determinar si el patrón de medidas postrespuesta difiere del patrón de medidas prerrespuesta. El uso de dicho análisis estadístico requiere la disponibilidad de muchos puntos de datos.

El análisis estadístico de medidas de tiempo representa un problema complicado y especial: las pruebas comunes de significancia que se aplican para medidas de tiempo pueden generar resultados falsos. Una razón de esto es que tales datos tienden a ser altamente variables y es fácil interpretar equívocamente cambios que no se deben a X, como si lo fueran. Es decir, con datos de tiempo las puntuaciones individuales y medias tienden a moverse bastante. Es fácil caer en la trampa de considerar uno de estos cambios como "significativo", en especial si va de acuerdo con la hipótesis. Si es posible suponer legítimamente que otras influencias, diferentes a X —ambas aleatorias y sistemáticas— son

uniformes sobre todas las series de Y, entonces el problema estadístico puede resolverse. Pero dicho supuesto puede ser, y con frecuencia es injustificado.

El investigador que explora estudios de tiempo debe estudiar con especial cuidado los problemas estadísticos y consultar a un especialista en estadística. Para el practicante la complejidad estadística resulta desafortunada, ya que quizá desmotive la realización de estudios prácticos necesarios. Puesto que los diseños longitudinales de un solo grupo se adaptan particularmente bien con la investigación de clase individual, se recomienda que en estudios longitudinales de métodos o estudios de niños en situaciones educativas, el análisis sea confinado al trazado de gráficas de resultados, y a su interpretación cualitativa. No obstante, las pruebas cruciales, especialmente aquellas para estudios publicables, deben reforzarse con pruebas estadísticas.

Diseño de series de tiempo múltiples

El diseño de series de tiempo múltiples es una extensión del diseño de series de tiempo interrumpidas. Con el diseño de series de tiempo interrumpidas solamente se utilizó un grupo de participantes; como resultado, las explicaciones alternativas pueden provenir de un efecto de la historia. El diseño de series de tiempo múltiples tiene la ventaja de que elimina el efecto de la historia al incluir un grupo control compuesto de un grupo equivalente de participantes —o por lo menos comparable— que no recibe la condición de tratamiento. Lo anterior se presenta en el diseño 22.3, donde un grupo experimental recibe la condición de tratamiento y el grupo control no. Como consecuencia, el diseño ofrece un mayor grado de control sobre las fuentes de explicaciones alternativas o hipótesis rivales. Los efectos de la historia, por ejemplo, se controlan debido a que ejercerían la misma influencia en el grupo experimental que en el grupo control.

Diseño 22.3: Un diseño de series de tiempo múltiples

Y_1	Y_2	Y_3	Y_4	X	Y_5	Y_6	Y_7	Y_8	(Experimental)
Y_1	Y_2	Y_3	Y_4		Y_5	Y_6	Y_7	Y_8	(Control)

Existen, naturalmente, otras variaciones posibles del diseño 22.2, además del diseño 22.3. Una variación importante consiste en añadir uno o más grupos control; otra sería añadir más observaciones de tiempo. Otra más consistiría en agregar más X, más intervenciones experimentales (véase Gottman, 1981; Gottman, McFall y Barnett, 1969; Campbell y Stanley, 1963).

Diseños experimentales de un solo sujeto

La mayoría de la investigación del comportamiento actual implica el uso de participantes. Sin embargo, existen otros métodos. En esta sección se analizan las estrategias para lograr control en los experimentos por medio del uso de uno o de unos pocos participantes. Estos diseños de un solo sujeto algunas veces se llaman diseños $N = 1$. Los diseños de un solo sujeto son una extensión del diseño de series de tiempo interrumpidas. Mientras las series de tiempo interrumpidas generalmente observan un grupo de individuos a través del tiempo (por ejemplo, niños), el estudio de un solo sujeto utiliza únicamente un participante o, cuando mucho, pocos participantes. Aun cuando se utilicen pocos participantes, cada uno es estudiado individual y extensamente; éstos también se llamarán

diseños o estudios de un solo sujeto. Aunque tengan diferentes nombres, todos comparten las siguientes características:

- Solamente se utilizan uno o pocos participantes en el estudio.
- Cada sujeto participa en varios ensayos (medidas repetidas), lo cual es similar a los diseños dentro de participantes que se describieron en el capítulo 21.
- Los procedimientos de aleatorización (por ejemplo, asignación aleatoria y/o selección aleatoria) se utilizan en muy raras ocasiones. En su lugar, las mediciones repetidas o intervalos de tiempo se asignan aleatoriamente a las diferentes condiciones de tratamiento.

Estos diseños observan el comportamiento del organismo antes del tratamiento experimental y utilizan las observaciones como una medida de la línea base. Las observaciones realizadas después del tratamiento se comparan, posteriormente, con las observaciones de la línea base; el participante sirve como su propio control. Estos diseños por lo común se aplican en investigación escolar, clínica y de asesoría. Se utilizan para evaluar los efectos de intervenciones conductuales a través del tiempo. Este tipo de investigación es popular entre quienes realizan experimentos sobre aprendizaje operante o modificación conductual.

La investigación con participantes únicos no es nueva, como lo ilustró Gustav Fechner, quien desarrolló la disciplina de la psicofísica en la década de 1860, utilizando solamente dos participantes: él y su cuñado. A Fechner se le acredita como el inventor de los métodos psicofísicos básicos que aun hoy en día se utilizan para medir los umbrales sensoriales. Fechner ejerció una fuerte influencia en Hermann Ebbinghaus, conocido por su trabajo experimental sobre la memoria. Ebbinghaus también se utilizó a sí mismo como sujeto. Wilhelm Wundt, considerado como el fundador del primer laboratorio psicológico en 1879, condujo experimentos donde medía varias respuestas psicológicas y de conducta en participantes individuales. I. P. Pavlov realizó su trabajo pionero sobre el condicionamiento instrumental utilizando perros de forma individual. La lista de psicólogos que han utilizado participantes únicos es extensa, la mayoría de los cuales lo hicieron antes de 1930 y antes del advenimiento del trabajo de R. A. Fisher y William Sealy Gossett sobre estadística moderna.

Los científicos del comportamiento que realizaron investigación antes del desarrollo de la estadística moderna intentaron resolver el problema de la confiabilidad y de la validez llevando a cabo extensas observaciones y réplica frecuentes de los resultados. Éste es un procedimiento tradicional utilizado por los investigadores que conducen experimentos con un solo sujeto. El supuesto es que los participantes individuales son, en esencia, equivalentes y que se requiere estudiar participantes adicionales tan sólo para asegurarse de que el sujeto original estaba dentro de la *norma*.

La popularidad del trabajo de Fisher sobre el análisis de varianza, y el estudio de Gossett sobre la prueba *t* de Student, abrieron el camino para la metodología de investigación orientada a grupos. Algunos afirman que estos trabajos fueron tan populares que la tradición del sujeto único casi se extinguió. De hecho, incluso en el mundo actual, existen criterios de contratación en las principales universidades, que dependen de si el candidato es un científico orientado hacia la investigación de grupos o un investigador orientado hacia los diseños de participantes únicos. A pesar de la popularidad de los métodos de Fisher y de la investigación orientada a grupos, algunos psicólogos continúan trabajando en la tradición del sujeto único. El más notable fue Burrus Frederick Skinner. Skinner se abstiene de utilizar estadística inferencial; no recomienda el uso de estadística inferencial compleja. En cambio considera que es posible demostrar la eficacia del tratamiento al graficar las acciones del comportamiento del organismo a través del tiempo; él llamó a

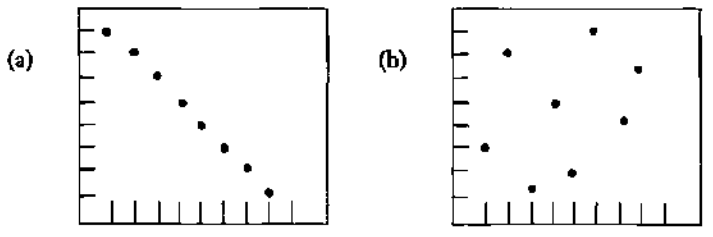
esto el registro acumulativo. Algunos, como E. L. Thorndike, le llaman "curva de aprendizaje". Skinner considera que es más útil estudiar un animal durante 1 000 horas, que estudiar 1 000 animales durante una hora. En su libro clásico, Sidman (1960) describe la filosofía de investigación de Skinner y hace una clara distinción entre el método de investigación de un solo sujeto y el método de investigación de grupo. El primero supone que la varianza del comportamiento del sujeto está dictada por la situación; como resultado, esta varianza puede eliminarse a través de un control experimental cuidadoso. El segundo método, en cambio, supone que la mayoría de la variabilidad es inherente y puede controlarse y analizarse estadísticamente.

Algunas ventajas de los estudios de un solo sujeto

La investigación orientada hacia grupos generalmente incluye el cálculo de la media o de alguna otra medida promedio o de tendencia central; pero los promedios pueden confundir. Observe los gráficos a) y b) en la figura 22.3 —ambos tienen exactamente los mismos valores—. Si se calculara la media de los datos de cada figura, se encontraría que son exactamente iguales. Incluso, si se calcularan la desviación estándar o la varianza, resultaría que las dos medidas de variación son exactamente las mismas. No obstante, una inspección visual de los datos muestra que el gráfico 22.3 a), exhibe una tendencia, mientras que el gráfico 22.3 b) no la muestra. De hecho, el gráfico 22.3 b) presenta lo que parece ser un patrón aleatorio. El método de un solo sujeto no tiene este problema, ya que se estudia a un participante de forma extensa a través del tiempo. El registro acumulativo de ese participante muestra el desempeño real de dicho participante.

Uno de los principales problemas del uso de muestras grandes es que la significancia estadística se logra por medio de diferencias muy pequeñas. Con la estadística inferencial, una muestra grande tenderá a reducir la cantidad de varianza del error. Tome la prueba *t* como ejemplo. Aun cuando la diferencia entre las medias permanezca igual, el incremento en el tamaño de la muestra tenderá a disminuir el error estándar. Con la reducción del error estándar, el valor de *t* se vuelve más grande, incrementando así su posibilidad de significancia estadística. Sin embargo, la significancia estadística y la significancia práctica son dos cuestiones diferentes. El experimento puede tener poca significancia práctica aun cuando tenga enorme significancia estadística. Simon (1987) criticó el uso indiscriminado de grupos grandes de participantes, pues los considera un desperdicio e incapaces de producir información útil. Simon recomienda el uso de experimentos de rastreo para encontrar las variables independientes con el mayor efecto sobre la variable dependiente; éstas serían las variables poderosas que producen grandes efectos. Simon no apoya exactamente los diseños de un solo sujeto; él recomienda el uso de diseños bien construidos, con el número necesario de participantes para encontrar los mayores efectos. Simon (1976) se refiere a esto como los "diseños económicos multifactoriales". Los investigadores de un solo sujeto, por otro lado, favorecen el incremento del tamaño del efecto en lugar de

FIGURA 22.3 *Comparación del grupo experimental y el grupo control*



intentar reducir la varianza del error, pues consideran que esto puede realizarse por medio de un control más rígido del experimento.

En la misma línea, los diseños de un solo sujeto tienen la ventaja, sobre los diseños orientados a grupos, de que con sólo unos cuantos participantes los investigadores pueden probar diferentes tratamientos. En otras palabras, determinan la eficacia o la no eficacia de una intervención de tratamiento sin emplear un número grande de participantes, lo cual puede resultar costoso.

Con los estudios de un solo sujeto, el investigador puede evitar algunos de los problemas éticos que enfrentan los investigadores orientados a grupos. Uno de dichos problemas éticos se refiere al grupo control, en el cual algunas situaciones no recibe ningún tratamiento real. Aunque en la mayoría de los estudios realizados hoy en día a los participantes del grupo control no se les daña en forma alguna, aún existen algunas cuestiones éticas. Considérese como ejemplo el estudio de Gould y Clum (1995) que buscaba determinar si la autoayuda, con mínimo contacto con el terapeuta, es efectiva en el tratamiento del trastorno de pánico. Todos los participantes de este estudio sufrían de ataques de pánico y fueron asignados aleatoriamente tanto al grupo experimental como al grupo control. El grupo experimental recibió material de autoayuda. El grupo control "no recibió tratamiento durante el transcurso del experimento" (p. 536). En su lugar, al grupo control se le indicó que estaban en la lista de espera para el tratamiento.

En el estudio de cierto tipo de individuos el tamaño de la población es pequeño y, por lo tanto, sería difícil adecuar el muestreo u obtener suficientes participantes para el estudio. De hecho, el estudio de Strube (1991) indica que incluso el muestreo aleatorio tiende a fallar cuando se utilizan muestras pequeñas. Si no se tienen disponibles suficientes participantes con ciertas características para el estudio, entonces el investigador puede considerar el uso de diseños de un solo sujeto, en lugar de abandonar el estudio. Simon (1987) cita el estudio que intentaron realizar Adelson y Williams, en 1954, sobre los parámetros de entrenamiento importantes en la educación de los pilotos. El estudio fue abandonado pues había demasiadas variables que considerar y no se tenían suficientes participantes. Simon señaló que el estudio pudo haberse realizado, pero no con el uso de la metodología tradicional orientada hacia grupos.

Algunas desventajas del diseño de un solo sujeto

Los estudios de un solo sujeto no están exentos de problemas o limitaciones. Algunas de ellas se harán más notorias cuando se analicen los tipos de diseños de un solo sujeto. Uno de los problemas más generales del paradigma de un solo sujeto es la validez externa. Algunos encuentran difícil creer que los hallazgos de un estudio que utilice un sujeto (o quizá tres o cuatro) puedan generalizarse a la población entera.

Con ensayos repetidos en un participante puede cuestionarse si el tratamiento sería igualmente eficaz para un participante que no ha experimentado tratamientos previos. Si se habla de un tratamiento terapéutico, entonces la eficacia quizá radique en la acumulación de sesiones, en lugar de una sola sesión. La persona que está en el ensayo *enésimo* puede ser una persona muy diferente de la que se encuentra en el primer ensayo. Es aquí donde la investigación orientada hacia grupos puede eliminar este problema; el tratamiento se aplica a cada persona solamente una vez.

Los estudios de un solo sujeto son quizás aún más sensibles a las aberraciones por parte del experimentador y del participante. Dichos estudios son eficaces sólo cuando el investigador puede evitar sesgos y el participante está motivado y coopera. El investigador puede mostrar una tendencia a observar sólo ciertos efectos y a ignorar otros. Se analizó el caso de Blondlot en este libro; él era el único científico capaz de ver los "rayos-N". No era tanto cuestión de que fuera un fraude, sino de que estaba sesgado hacia ver algo que no

estaba ahí. Un investigador que hace investigación con un solo sujeto se ve afectado más de esta forma que el investigador orientado a grupos, y requiere desarrollar un sistema de verificación y balances para evitar esta dificultad.

Numerosas investigaciones requieren, por naturaleza, seguir métodos orientados a grupos y, como tales, serían impropios para diseños de un solo sujeto. Por ejemplo, para estudiar el comportamiento de miembros del jurado se requeriría el uso de grupos y la influencia de la dinámica de grupos. Antes se analizó la investigación alrededor del pensamiento grupal. El estudio de este importante fenómeno fue mejor realizado con grupos, ya que fue el grupo como un todo el que mostró dicho fenómeno.

Algunos paradigmas de la investigación de un solo sujeto

La línea base estable: una meta importante

En un diseño orientado hacia grupos, un grupo de participantes se compara con otro; o se compara un grupo de participantes que recibe una condición con el mismo conjunto de participantes que reciben una condición diferente. Se supone que los grupos son iguales antes de la aplicación del tratamiento de tal manera que, si la variable dependiente difiere después del tratamiento, se pueden asociar tales diferencias con el tratamiento. La determinación de un tratamiento eficaz se realiza al comparar estadísticamente la diferencia entre los dos grupos respecto a alguna variable resultante. Sin embargo, cuando se utiliza un solo sujeto, debe emplearse una táctica diferente. En la situación de un sujeto es necesario comparar el comportamiento que ocurre antes con el comportamiento que ocurre después de la introducción de una intervención experimental. El comportamiento previo a la intervención del tratamiento debe medirse durante un periodo lo suficientemente grande para poder obtener una línea base estable. Dicha línea base o nivel operante es importante porque se compara con el comportamiento posterior. Si la línea base varía de manera considerable, entonces sería más difícil evaluar cualquier cambio confiable en el comportamiento después de la intervención. El problema de la línea base con los diseños de un solo sujeto es importante. Para encontrar una descripción completa del problema y de sus posibles soluciones, se debe consultar a Barlow y Hersen (1984). Otra excelente referencia es Kazdin (1982).

Un ejemplo donde las medidas de línea base son muy importantes es el uso de un polígrafo (detector de mentiras). Aquí, el operador obtiene mediciones fisiológicas de la persona (sospechoso). El operador formula ciertas preguntas al sospechoso, cuya respuesta se sabe cierta (nombre, color de ojos, lugar de nacimiento, etcétera). Las respuestas emitidas se registran y se utilizan como medida de línea base para las respuestas honestas. Se toma otra línea base para respuestas conocidas como falsas: se le indica al sospechoso mentir deliberadamente a las preguntas planteadas. Después del establecimiento de estas dos líneas base, se plantea la pregunta de importancia (v.g., ¿cometió usted el crimen?) y se compara con las dos líneas base. Si la respuesta en el polígrafo se asemeja a la línea base de mentir, entonces se considera que el sospechoso mintió.

Diseños que utilizan el retiro del tratamiento

El diseño ABA

El diseño ABA incluye tres grandes pasos. El primero consiste en establecer una línea base estable (A). En el segundo paso (B) se aplica la intervención experimental al participante.

Si el tratamiento es efectivo, habrá una respuesta diferente a la de la línea base. Para determinar si la intervención del tratamiento causó el cambio en el comportamiento, el investigador lleva a cabo el paso tres: un regreso a la línea base (A). El tercer paso se requiere porque no se sabe cuál habría sido la tasa de respuesta si el participante no recibiera tratamiento. También se necesita saber si el cambio en la respuesta se debió a la intervención del tratamiento o a algo más.

Un problema importante del diseño ABA es que el efecto de la intervención puede no ser completamente reversible. Si el tratamiento implicó una cirugía, donde se removió el hipotálamo o se seccionó el cuerpo calloso, sería imposible revertir estos procedimientos. Un método de aprendizaje que provoque algún cambio permanente en el comportamiento del participante no sería reversible.

Existen también algunas consideraciones éticas respecto a regresar al paciente al estado original, si tal estado fuese un comportamiento indeseable (Tingstrom, 1996). Los experimentos en modificación conductual rara vez regresan al participante a la línea base. Este regreso a la línea base se llama condición de retiro. Para beneficiar a los participantes, se reintroduce el tratamiento. El diseño ABAB hace esto.

Repetición de tratamientos (diseño ABAB)

Existen dos versiones del diseño ABAB. El primero se describió brevemente en la sección anterior. El diseño ABAB es igual al diseño ABA, excepto que el tratamiento se reintroduce al participante y éste deja el estudio después de lograr cierto nivel benéfico. La repetición del tratamiento también proporciona al experimentador información adicional sobre la fortaleza de la intervención del tratamiento. El hecho de demostrar que la intervención del tratamiento puede llevar al participante al nivel de beneficio previo, después de regresar a la persona a la línea base, da fuerza a la afirmación de que el tratamiento causó el cambio en el comportamiento; es decir, brinda evidencia de validez interna. El diseño ABAB esencialmente produce el efecto experimental dos veces.

La segunda variante del diseño ABAB es el llamado diseño de *tratamientos alternantes*. En esta variante no se toma la línea base. A y B en este diseño son dos tratamientos diferentes que se alternan aleatoriamente. El objetivo de este diseño consiste en evaluar la eficacia relativa de las dos intervenciones de tratamiento. A y B pueden ser dos métodos diferentes para controlar la alimentación en exceso. Cada tratamiento se aplica al participante en diferentes momentos. Después de un periodo, un método puede emerger como más efectivo que el otro. La ventaja que tiene este diseño sobre el primer diseño ABAB es que no se requiere obtener una línea base y el participante no está sujeto a procedimientos de retiro. Puesto que este método implica la comparación de dos conjuntos de series de datos, algunos lo llaman diseño entre series.

Existen otras variantes interesantes del diseño ABAB, donde no se lleva a cabo el retiro del tratamiento. McGuigan (1996) lo llama el diseño ABCB. En la tercera fase de este diseño, el organismo recibe una condición "placebo". La condición placebo es esencialmente un método diferente.

Los diseños de un solo sujeto se diferencian de los diseños de grupo en que sólo permiten que el investigador varíe una variable a la vez. El investigador no sería capaz de determinar qué variable o qué combinación de ellas causó los cambios en la respuesta, si dos o más variables se alteraron simultáneamente. Lo mejor que cualquiera puede hacer es afirmar que la combinación de las variables condujo al cambio. Sin embargo, el investigador será incapaz de determinar cuál o qué tanto de cada una; si hay dos variables, llamadas B y C, y la línea base es A, entonces una posible secuencia de presentación de las condiciones sería A-B-A-B-BC-B-BC. En dicha secuencia cada condición es precedida y procedida por la misma condición una vez por lo menos, con una sola variable cambiando a la vez.

El diseño A-B-A-B-BC-B-BC con frecuencia se denomina un diseño de interacción. Sin embargo, no están presentes todas las combinaciones posibles de B y C. La condición C nunca ocurre sola (A representa la ausencia de B y C). Esta interacción difiere de las interacciones analizadas en el capítulo sobre diseños factoriales. Lo que se prueba con este procedimiento es si C se añade o no al efecto de B.

En un experimento de aprendizaje que utilice este diseño se podría examinar el efecto de elogiar a un estudiante por dar la respuesta correcta (C) a una pregunta sobre geografía, aunado a un punto meritorio (B). Si se descubre que el elogio, junto con el punto meritorio tienen un mayor efecto que el punto meritorio solo, se tiene información útil para diseñar una situación de aprendizaje para éste y otros estudiantes; pero no se conocerá el efecto singular del elogio. Utilizar únicamente el elogio podría haber sido tan efectivo como el punto meritorio más el elogio; o quizá emplear tan sólo el elogio hubiera tenido poco o ningún efecto. Sin embargo, es posible evaluar el elogio extendiendo el diseño de un solo sujeto: la secuencia A-B-A-B-BC-B-BC-C-BC. Pero extender un experimento de un solo sujeto de este tipo acarrea otros problemas; por ejemplo, un sujeto puede experimentar fatiga o perder el interés. Como resultado, una sesión demasiado larga quizá no produzca información útil, aunque el diseño parezca correcto.

Un ejemplo de investigación

Powell y Nelson: ejemplo de un diseño ABAB

Este estudio de Powell y Nelson (1997) incluyó un participante, Evan, un niño de 7 años de edad, diagnosticado con trastorno por déficit de atención con hiperactividad (TDAH). Evan recibía 15 mg de Ritalin® al día. La mayor parte de su comportamiento en el salón de clases se consideraba indeseable; también tenía relaciones pobres con sus compañeros y no comprendía su trabajo escolar. Las conductas indeseables incluían falta de obediencia, abandono de su escritorio, molestar a otros, iniciar las actividades a destiempo y no realizar su trabajo. Los datos se recolectaron a través de las interacciones entre Evan y su maestro.

El tratamiento consistió en permitir a Evan elegir las materias en las cuales deseaba trabajar. Había dos condiciones: elección y no elección. Los datos de la línea base fueron recolectados durante la fase de no elección; a Evan se le impartió la misma materia que al resto de la clase. Durante las fases de elección el maestro le presentó a Evan tres materias diferentes y él eligió una. Las opciones eran materias idénticas en longitud y dificultad y sólo variaban respecto a su contenido. A Evan no se le dio la misma opción de materias dos veces.

Powell y Nelson utilizaron un diseño ABAB para evaluar los efectos de la toma de decisiones sobre la conducta indeseable de Evan. Los resultados mostraron que durante la condición de elección disminuyó el número de conductas indeseables. Dicho estudio apoyó la eficacia de la toma de decisiones como técnica control de antecedente. Tales resultados sugieren que los educadores que intentan manejar la conducta de los estudiantes en clase pueden utilizar procedimientos de elección.

Uso de líneas base múltiples

Existe una forma de investigación de un solo sujeto que emplea más de una línea base. Se establecen varias líneas base antes de aplicar el tratamiento al participante. Estos tipos de estudios se llaman *estudios de líneas base múltiples*. Existen tres clases de diseños de investigación de líneas base múltiples: a través de conductas, a través de participantes y a través de escenarios.

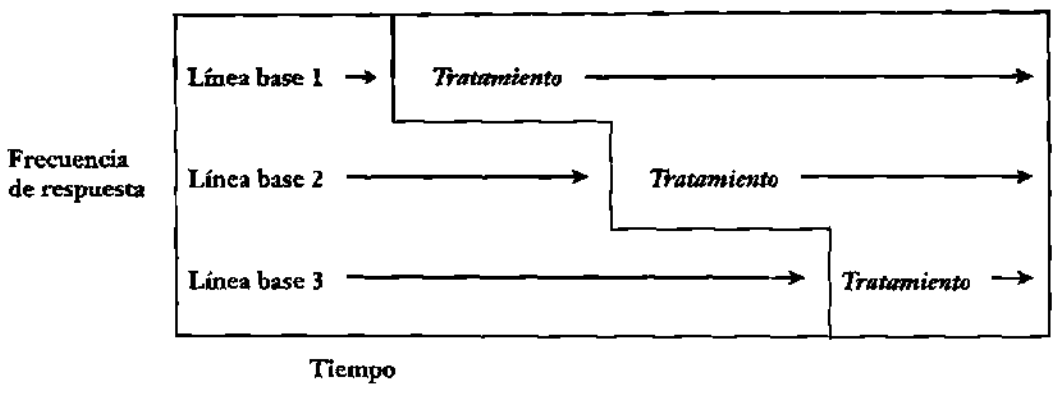
El uso de líneas base múltiples constituye otro método para demostrar la eficacia de un tratamiento sobre el cambio en el comportamiento. Existe un patrón común para la implementación de las tres clases de este diseño, cuyo patrón se muestra en la figura 22.4.

Con las líneas base múltiples *a través de conductas*, las intervenciones de tratamiento para cada conducta diferente se introducen en diferentes momentos. En la figura 22.4, cada línea base sería de una conducta diferente. En el caso de un niño autista, la línea base 1 podría ser golpear su cabeza contra la pared. La línea base 2 tal vez sea hablar constantemente en diferentes tonos añadiendo ruidos. La línea base 3 sería golpear a otros. Se establecen las tres líneas base para saber si el cambio en las conductas coincide con la intervención del tratamiento. Si una de las conductas cambia, mientras las otras permanecen constantes o estables en la línea base, el investigador podría afirmar que el tratamiento resultó eficaz con esa conducta específica. Después de que pase cierto periodo, se aplica el mismo tratamiento a la segunda conducta indeseable (línea base 2). Cada conducta subsecuente se somete al tratamiento en el mismo procedimiento paso a paso. Si la intervención del tratamiento es eficaz al cambiar la tasa de respuesta de cada conducta, entonces es posible afirmar que el tratamiento fue eficaz.

Una consideración importante con esta clase particular de diseño de líneas base múltiples, es que parte del supuesto de que las respuestas de cada conducta son independientes de las respuestas de otras conductas. La intervención puede considerarse eficaz si existe tal independencia. Si las respuestas están correlacionadas de alguna manera, entonces la interpretación de los resultados se torna más difícil.

En el diseño de líneas base múltiples *a través de participantes* se aplica el mismo tratamiento en series a la misma conducta de distintos individuos en el *mismo ambiente*. Ahora cada línea base en la figura 22.4 representa un participante diferente. Cada participante recibe el mismo tratamiento para la misma conducta, en el mismo ambiente. El estudio de Tingstrom, Marlow, Edwards, Kelshaw y Olmi (1997) constituye un ejemplo de un estudio de líneas base múltiples a través de participantes. El paquete de entrenamiento de obediencia es la intervención del tratamiento. La intervención utiliza *tiempo dentro* (contacto físico y elogio verbal) y *tiempo fuera* (un procedimiento coercitivo) para aumentar la tasa de obediencia del estudiante hacia las instrucciones del maestro. La conducta de interés es la obediencia a las instrucciones del maestro. El ambiente es el salón de clases. Los participantes de este estudio fueron tres estudiantes —A, B y C— quienes habían mostrado conducta de no obediencia. Los tres estudiantes presentaban trastornos de articulación

FIGURA 22.4 *Formato general del diseño de líneas base múltiples*



y de lenguaje. El diseño del estudio se adhirió a las siguientes fases de intervención: línea base, sólo tiempo dentro, tiempo dentro y tiempo fuera combinados, y seguimiento. Los estudiantes B y C permanecieron en la línea base mientras se implementaba la fase de sólo tiempo dentro para el estudiante A. Cuando A mostró un cambio en la obediencia, se implementó la fase sólo tiempo dentro para B, mientras que C permanecía en línea base. Cuando B mostró un cambio en la obediencia, se implementó sólo tiempo dentro para C. Tingstrom y sus colaboradores fueron capaces de demostrar la efectividad de la intervención combinada de tiempo dentro y tiempo fuera para incrementar la obediencia.

En el diseño de líneas base múltiples *a través de escenarios*, el mismo tratamiento se aplica a diferentes participantes, quienes se encuentran en *diferentes escenarios*. En este diseño, cada línea base en la figura 22.4 representa a un participante diferente en un ambiente diferente. El tratamiento y la conducta bajo estudio serían los mismos. Aquí es posible tener tres pacientes diferentes, cada uno residente de un tipo distinto de instalación de cuidado psiquiátrico, como las estudiadas por Nelson y sus colaboradores, que se analizaron previamente en este capítulo. En dicho estudio Nelson, Hall y Walsh-Bowers (1997) compararon los efectos a largo plazo de apartamentos de apoyo (AA), hogares de grupo (HG) y hogares de alojamiento y cuidado (HAC).

RESUMEN DEL CAPÍTULO

1. Los experimentos verdaderos son aquellos en que el experimentador puede seleccionar a los participantes de manera aleatoria, asignar a los participantes a las condiciones de tratamiento aleatoriamente y controlar la manipulación de la variable independiente. El diseño cuasi-experimental carece de una o más de estas características.
2. Cook y Campbell (1979) analizan ocho variantes del diseño de grupo control no equivalente. El que cubre este libro es el diseño de grupo control sin tratamiento. Se analizan cinco resultados diferentes en términos de la validez interna.
3. Los diseños de series de tiempo son diseños longitudinales que incluyen mediciones repetidas de las mismas variables dependientes en diferentes intervalos de tiempo fijos. Casi siempre en cierto momento se introduce la intervención del tratamiento.
4. La selección y la interacción entre la selección y la maduración son dos explicaciones alternativas que cubren los resultados obtenidos a partir de los diseños cuasi-experimentales.
5. Los experimentos que utilizan participantes únicos no son nuevos. Los investigadores pioneros de la psicología experimental utilizaron diseños de un solo sujeto.
6. Los investigadores consideran que en los diseños de un solo sujeto la variabilidad de la situación se elimina con el control experimental adecuado.
7. Los investigadores orientados a grupos consideran que la variabilidad puede analizarse estadísticamente.
8. La investigación de un solo sujeto posee varias ventajas sobre la investigación de grupos, en términos de flexibilidad y ética. Sin embargo, carece de credibilidad de la validez externa.
9. Los efectos pequeños, pero estadísticamente significativos, encontrados en la investigación de grupos pueden tener poca significancia clínica o práctica, y haberse introducido artificialmente por tamaños grandes de muestras. Cuando esto sucede, el tamaño del efecto es pequeño. La investigación de un solo sujeto se concentra en el tamaño del efecto y no en el tamaño de la muestra.

10. El establecimiento de una línea base estable constituye una de las tareas más importantes en la investigación de un solo sujeto.
11. Al establecimiento de una línea base, seguido por la administración de un tratamiento y, después, por el retiro del tratamiento, se le llama diseño ABA.
12. Un problema importante del diseño ABA es que tal vez el tratamiento sea irreversible, dejando al participante en el estado de mejoría, en lugar de regresar a la persona al estado indeseable original.
13. Una variación del diseño ABA es el diseño ABAB, donde se reinstaura el estado de mejoría del participante.
14. En un estudio de un solo sujeto se puede variar únicamente una variable independiente a la vez.
15. El llamado diseño de interacción no permite probar una interacción del tipo definido previamente en los diseños factoriales. Solamente examina dos variables en conjunto.
16. Existen tres tipos de diseños de líneas base múltiples. En cada caso, la intervención se introduce en diferentes momentos para diferentes conductas, participantes o escenarios. Si los cambios en las conductas coinciden con la introducción del tratamiento, esto proporciona evidencia de que el tratamiento es eficaz.

SUGERENCIAS DE ESTUDIO

1. Revise cada uno de los siguientes estudios y determine cuáles son diseños cuasiexperimentales, de grupo control no equivalente y de un solo sujeto.

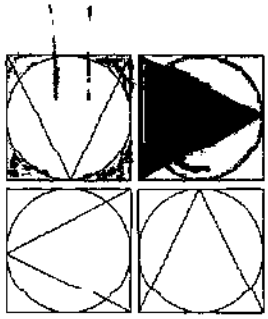
   Adkins, V. K. y Matthews, R. M. (1997). Prompted voiding to reduce incontinence in community-dwelling older adults. *Journal of Applied Behavior Analysis, 30,* 153–156.

   Lee, M. J. y Tingstrom, D. H. (1994). A group math intervention: The modification of cover, copy, and compare for group application. *Psychology in the Schools, 31,* 133–145.

   Streufert, S., Satish, U., Pogash, R., Roache, J. y Severs, W. (1997). Excess coffee consumption in simulated complex work settings: Detriment or facilitation of performance? *Journal of Applied Psychology, 82,* 774–782.

2. ¿Por qué es necesaria una medida de línea base en los diseños de un solo sujeto?
3. ¿Los datos de diseños de un solo sujeto deben analizarse estadísticamente? Explique por qué.
4. Dé un ejemplo donde deba utilizarse un diseño de un solo sujeto. También cite una situación de investigación donde resulte más apropiado un diseño de grupo.
5. Un estudiante universitario desea realizar un estudio de series de tiempo sobre los efectos de la luna llena en pacientes psiquiátricos. ¿Qué variable dependiente debe utilizar? ¿Dónde debe buscar para obtener datos para un estudio de este tipo?
6. ¿Los diseños de un solo sujeto son aplicables a la investigación médica? ¿Deben enseñarse diseños de un solo sujeto a los estudiantes de escuelas de medicina? Lea el siguiente artículo:

   Bryson-Brockmann, W. y Roll, D. (1996). Single-case experimental designs in medical education: An innovative research method. *Academic Medicine, 71,* 78-85.



CAPÍTULO 23

INVESTIGACIÓN NO EXPERIMENTAL

Entre las falacias prevalecientes, una de las más peligrosas para la ciencia es la conocida como *post hoc, ergo propter hoc*: después de esto, por lo tanto, causado por esto. Es posible bromear al decir con un toque de seriedad, "si me llevo mi paraguas, no lloverá". Incluso se podría decir seriamente que los delincuentes son delincuentes debido a la falta de disciplina en las escuelas, o que la educación religiosa hace a los niños más virtuosos. Es fácil asumir que una cuestión causa otra, simplemente porque ocurre antes de la otra y porque se tiene una amplia gama de "causas" posibles. Entonces, también muchas explicaciones parecen frecuentemente plausibles. Por ejemplo, es fácil creer que el aprendizaje de los niños mejora porque se instituye una nueva práctica educativa o porque se enseña de cierta manera. Se asume que la mejoría en su aprendizaje se debió al nuevo

método de ortografía, a la institución de los procesos de grupo en la situación dentro del salón de clases, a una disciplina severa y a mayor cantidad de tarea (o poca disciplina y menos tarea). En raras ocasiones se considera que los niños aprenderán algo si se les da la oportunidad de aprender.

El científico social y el científico educativo enfrentan a menudo el problema de la falacia *post hoc*. El sociólogo que busca las causas de la delincuencia sabe que debe tenerse extremo cuidado al estudiar el problema. Las condiciones de pobreza, los hogares destruidos, las cantidades de plomo en las tuberías de agua, la carencia de amor —todas y cada una— son causas posibles de la delincuencia. El psicólogo que busca las raíces de la personalidad adulta enfrenta un problema aún más sutil: rasgos heredados, prácticas de crianza, influencias educativas, personalidad de los padres y circunstancias ambientales; todas son explicaciones plausibles. El científico educativo, con la meta de entender las bases de logro del éxito escolar, enfrenta también un gran número de posibilidades razonables: inteligencia, aptitud, motivación, ambiente familiar, personalidad del maestro, personalidad del alumno y métodos de enseñanza.

El peligro del supuesto *post hoc* es que puede, y con frecuencia lo hace, conducir a interpretaciones erróneas y confusas de los datos de investigación, cuyo efecto es particularmente serio cuando los científicos tienen poco o ningún control sobre el tiempo y sobre las variables independientes. Cuando buscan explicar un fenómeno que ya ha ocurrido, los científicos se ven confrontados con el desagradable hecho de que no tienen un control real de las causas posibles. Por lo tanto, deben elaborar un curso de acción de investigación, que difiera en ejecución e interpretación del que siguen los científicos que realizan experimentación.

Definición

La investigación no experimental es la búsqueda empírica y sistemática en la que el científico no posee control directo de las variables independientes, debido a que sus manifestaciones ya han ocurrido o a que son inherentemente no manipulables. Se hacen inferencias sobre las relaciones entre las variables, sin intervención directa, de la variación concomitante de las variables independiente y dependiente.

Suponga que un investigador tiene interés en la relación del sexo con la creatividad de los niños. Mide la creatividad de una muestra de niños y niñas, y prueba la significancia de la diferencia entre las medias de los dos sexos. La media de los niños es significativamente mayor que la media de las niñas. Una conclusión es que los niños son más creativos que las niñas y quizá no sea una conclusión válida. La relación existe, es verdad; sin embargo, con esta sola evidencia la conclusión resulta dudosa. Surgiría una pregunta: ¿es la relación demostrada realmente entre el sexo y la creatividad? Como existen otras variables que están correlacionadas con el sexo, una o más de ellas pudieron haber generado la diferencia entre las puntuaciones de creatividad de los dos sexos.

Diferencia básica entre la investigación experimental y la no experimental

La base de la estructura sobre la que la ciencia experimental opera es simple. Se hipotetiza: si x, entonces y; si frustración, entonces agresión. Dependiendo de las circunstancias y de la predilección personal en el diseño de investigación, se utiliza un método para manipular o medir x. Entonces se observa y para observar si ocurre una variación concomitante, es

decir, la variación esperada o que se predice a partir de la variación de x. Si esto sucede, es evidencia de la validez de la proposición $x \rightarrow y$, "si x entonces y". Considere que aquí se predice y a partir de una x controlada. Para lograr mayor control, se puede utilizar el principio de aleatorización y manipulación activa de x, y se puede asumir que, siendo lo demás igual, y varía como resultado de la manipulación de x.

Por otra parte, en la investigación no experimental se observa y, y también se observan una o varias x. Éstas se observan ya sea antes, después o concomitantemente a la observación de y. No hay diferencia en la lógica básica. Es posible demostrar que la estructura del argumento y su validez *lógica* son iguales en la investigación experimental y en la no experimental. Además, el propósito básico de ambas es el mismo: establecer la validez *empírica* de las llamadas proposiciones condicionales de la forma: si p, entonces q. La diferencia esencial es el control directo de p, la variable independiente. En la investigación experimental p puede manipularse, lo cual es más bien un "control" directo. Cuando Clark y Walberg (1968) pidieron a unos maestros que dieran reforzamiento masivo a un grupo de participantes y a otros maestros que dieran reforzamiento moderado a otro grupo, estaban manipulando o controlando directamente la variable reforzamiento. De la misma forma, cuando Dolinski y Nawrat (1998) sometieron a un grupo a estrés (ansiedad), sometieron a otro grupo a estrés y después lo redujeron, e incluyeron un tercer grupo con poco o ningún estrés, estaban manipulando directamente la variable ansiedad. Además, los participantes pueden asignarse aleatoriamente a los grupos experimentales.

En la investigación no experimental el control *directo* no es posible: tampoco puede utilizarse la manipulación experimental ni la asignación aleatoria. Existen dos diferencias esenciales entre los modelos experimental y no experimental. A causa de la carencia de control relativo de x y de otras posibles x, la "verdad" de la relación hipotetizada entre x y y no puede sostenerse con la misma confianza que en la situación experimental. Básicamente la investigación no experimental tiene, por así decirlo, una debilidad inherente: la carencia de control de las variables independientes.

La diferencia más importante entre la investigación experimental y la investigación no experimental es, entonces, el *control*. En los experimentos, los investigadores tienen, por lo menos, control manipulativo: por lo menos tienen una variable activa. Si un experimento es un "verdadero" experimento, también puede ejercerse control por medio de la aleatorización. Es posible asignar participantes aleatoriamente a los grupos, o asignar los tratamientos aleatoriamente a los grupos. En la situación de investigación no experimental, este tipo de control de las variables independientes no es posible. Los investigadores deben tomar las cosas como son e intentar entenderlas.

Considere un caso bien conocido. Cuando se pinta la piel de las ratas con sustancias cancerígenas (x), se controlan adecuadamente otras variables, y las ratas finalmente desarrollan un carcinoma (y), el argumento es contundente, ya que se está controlando x (y otras x teóricamente posibles), y se predice y. Cuando se encuentran casos de cáncer pulmonar (y) y se regresa a revisar en la posible multiplicidad de causas ($x_1, x_2, \ldots, x_n$), y se elige el hábito de fumar cigarrillos (digamos x_3) como la causa, se está en una situación más difícil y ambigua. Ninguna situación resulta segura, por supuesto; ambas son probabilísticas. Pero en el caso experimental se puede estar considerablemente más seguro, si se mantuvieron "constantes el resto de las condiciones", de que la proposición si x, entonces y, es válida empíricamente. Sin embargo, en el caso no experimental no se pisa tierra tan firme, ya que no puede afirmarse con mucha certeza, "que se mantuvieron constantes el resto de las condiciones". No es posible controlar las variables independientes por medio de la manipulación o de la aleatorización. En pocas palabras, la probabilidad de que x esté "realmente" relacionada con y es mayor en la situación experimental que en la situación no experimental porque el control de x es mayor.

Autoselección e investigación no experimental

En un mundo ideal de la investigación del comportamiento, la obtención de muestras aleatorias de participantes, así como la asignación aleatoria de participantes y tratamientos a los grupos, siempre sería posible. No obstante, en el mundo real, ni una, ni dos e incluso ninguna de estas tres posibilidades existe. Es posible seleccionar participantes al azar, tanto en la investigación experimental como en la no experimental. Pero no es posible, en la investigación no experimental, asignar a los participantes aleatoriamente a los grupos o asignar los tratamientos aleatoriamente a los grupos. Los participantes pueden "asignarse a sí mismos" a los grupos. Es posible que se "seleccionen a sí mismos" en los grupos con base en características diferentes de aquellas que interesan al investigador. Los participantes y los tratamientos llegan como si ya hubieran sido asignados a los grupos.

La *autoselección* ocurre cuando los miembros de los grupos estudiados están en los grupos, en parte, porque poseen rasgos o características diferentes, ajenas al problema de investigación; características que posiblemente influyan o estén relacionadas con las variables del problema de investigación. Algunos ejemplos de autoselección ayudarán a una mejor comprensión.

En la bien conocida investigación sobre el tabaquismo y el cáncer, se estudiaron los hábitos de fumar de un gran número de personas. Este gran grupo se dividió en aquellos que padecían cáncer pulmonar —o que habían muerto por su causa— y aquellos que no lo padecían. Por lo tanto, la variable dependiente era la presencia o la ausencia del cáncer. Los investigadores exploraron los antecedentes de los participantes para determinar si fumaban cigarrillos, y si así era, cuántos. Fumar cigarrillos era la variable independiente. Los investigadores encontraron que la incidencia del cáncer pulmonar se incrementaba de acuerdo con el número de cigarrillos fumados por día. También encontraron que la incidencia fue menor en el caso de los fumadores moderados y de los no fumadores. Llegaron a la conclusión de que el hábito de fumar cigarrillos "causa" cáncer pulmonar. Esta conclusión puede o no resultar cierta; pero los investigadores no pueden llegar a esta conclusión, aunque afirmen que existe una relación estadísticamente significativa entre las variables. Observe que los investigadores científicos cuidadosos generalmente no utilizan el término *causa* a menos que el estudio haya sido realizado bajo las más estrictas condiciones. La palabra *causa* se utiliza aquí para ilustrar cómo los medios de comunicación masiva a menudo interpretan los hallazgos científicos que sugieren causalidad.

Los investigadores científicos tampoco pueden establecer una conexión causal porque existen muchas otras variables; y una de ellas o alguna combinación de ellas, que pudo causar el cáncer. Además ellos no controlaron otras variables independientes posibles. *No pueden* controlarlas excepto por medio de la comprobación de hipótesis alternativas, un procedimiento que se explicará más adelante. Aun cuando ellos también estudian "grupos control" de personas que no padecen cáncer, quizá esté operando la autoselección. Tal vez, por ejemplo, los hombres tensos y ansiosos están condenados a padecer cáncer pulmonar si se casan con mujeres altas. Puede suceder que este tipo de hombre también fume gran cantidad de cigarrillos. No es el tabaquismo lo que lo mata —él se mata a sí mismo al estar tenso y ansioso— y posiblemente por el hecho de casarse con una mujer alta. Tales hombres son seleccionados para formar parte de la muestra por los investigadores sólo porque fuman cigarrillos; sin embargo, ellos se seleccionan a sí mismos para la muestra porque comúnmente poseen un temperamento concomitante al tabaquismo.

La autoselección constituye un aspecto sutil. Existen dos tipos: 1) autoselección para *muestras* y 2) autoselección para *grupos comparativos*. Esto último sucede cuando se selecciona a los participantes porque pertenecen a un grupo o a otro: cáncer y sin cáncer, universitario y no universitario, bajo rendimiento o sin bajo rendimiento. Es decir, son

seleccionados *porque* poseen la variable dependiente en mayor o menor grado. La autoselección para muestras ocurre cuando se selecciona a los participantes de forma no aleatoria para la muestra.

Lo esencial del tema es que cuando la *asignación* no es aleatoria, siempre existe un resquicio para que otras variables se inmiscuyan. Cuando se colocan participantes dentro de los grupos, en el caso anterior o en casos similares, o cuando se "colocan ellos mismos" dentro de los grupos, con base en una variable, es posible que otra variable (o variables) correlacionada(s) con esta variable, sean la base "real" de la relación. El estudio no experimental común utiliza grupos que muestran diferencias respecto a la variable dependiente. En ciertos estudios de tipo longitudinal, los grupos se diferencian primero con base en la variable independiente. Pero ambos casos son básicamente iguales, ya que la pertenencia al grupo, con base en una *variable*, siempre conlleva la selección.

Por ejemplo, es posible seleccionar aleatoriamente a universitarios de primer año y después seguirlos para determinar la relación entre inteligencia y el éxito en la universidad. Los estudiantes se seleccionaron a sí mismos dentro de la universidad, por así decirlo. Una o más de las características que llevan consigo a la universidad, además de la inteligencia —nivel socioeconómico, motivación, antecedentes familiares— pueden ser los principales determinantes del éxito universitario. El hecho de iniciar con la variable independiente, en este caso la inteligencia, no cambia la naturaleza autoselectiva de la situación de investigación. En términos de muestreo, los estudiantes se seleccionaron a sí mismos para la universidad, lo cual sería un factor importante si se estudiaran alumnos universitarios y no universitarios. Pero si el interés radica tal sólo en el éxito y en el no éxito de estudiantes *universitarios*, entonces la autoselección para la universidad se vuelve irrelevante; mientras que la autoselección para los grupos de éxito y de no éxito resulta crucial. El hecho de medir la inteligencia de los estudiantes al entrar a la universidad, y seguirlos hacia el éxito y hacia el no éxito, no cambia ni el problema de selección ni el carácter no experimental de la investigación. En suma, los estudiantes se seleccionaron a sí mismos dentro de la universidad y se seleccionaron a sí mismos para tener o no éxito en la universidad.

Investigación no experimental a gran escala

Como siempre, los ejemplos de investigación ayudan a comprender la naturaleza de la investigación no experimental. En lugar de resumir únicamente estudios individuales, como se ha hecho hasta ahora, se describirán tanto estudios individuales como conjuntos de estudios centrados en algún fenómeno o variable de interés. La investigación no experimental del comportamiento con frecuencia se enfoca en grandes problemas de importancia social y humana: clase social, procesos políticos, segregación y disgregación, actitudes públicas, rendimiento escolar, por ejemplo. La importancia —*relevancia* es la palabra de moda— del tema de dichos estudios no debe oscurecer la comprensión de su carácter no experimental. Sin embargo, aunque la investigación no experimental tiene debilidades inherentes, eso no significa que la investigación experimental sea más importante. Como se mencionó antes, el experimento es uno de los grandes inventos de todos los tiempos, un ideal de control al que todos aspiran. Ello no significa que los experimentos sean necesariamente "mejores" que los estudios no experimentales. Por otro lado, la investigación no experimental no siempre es "mejor" que la investigación experimental porque su contenido y variables parezcan ser socialmente importantes. ¡Esto sería como decir que la investigación psicológica es "mejor" que la investigación sociológica, debido a que los psicólogos utilizan con mayor frecuencia un modelo experimental y los sociólogos un modelo no experimental!

Determinantes del rendimiento escolar

Una gran preocupación de los investigadores educativos ha sido la búsqueda de los factores determinantes del rendimiento escolar. ¿Qué factores conducen al éxito del rendimiento en la escuela? La inteligencia es un factor importante, por supuesto. Mientras que la inteligencia medida, especialmente la habilidad verbal, explica una gran proporción de la varianza del rendimiento, existen muchas otras variables, tanto psicológicas como sociológicas: sexo, raza, clase social, aptitud, características ambientales, condiciones de la escuela y habilidades del maestro, antecedentes familiares, métodos de enseñanza. El estudio del rendimiento está caracterizado tanto por modelos experimentales como no experimentales. Aquí nos ocupamos de estos últimos, ya que ilustran claramente los problemas de la investigación no experimental.

En 1966 se publicó el ahora famoso reporte Coleman (Coleman, Campbell, Hobson, McPartland, Mood, Weinfeld y York, 1966). Como su título indica (*Equality of Educational Opportunity*), fue un intento a gran escala para responder la pregunta: ¿ofrecen las escuelas estadounidenses iguales oportunidades educativas a todos los niños? Sin embargo, de igual importancia era la pregunta sobre la relación entre el rendimiento de los estudiantes y el tipo de escuela a la que asistían. Dicho estudio representó un esfuerzo masivo y admirable por responder estas preguntas (y otras). Su hallazgo más famoso y controvertido fue que las diferencias entre las escuelas explican sólo una pequeña fracción de las diferencias en el rendimiento escolar. La mayoría de la varianza del rendimiento estaba explicada por aquello que los niños llevaban consigo a la escuela. Hubo mucho que cuestionar respecto a la metodología y conclusiones del estudio. De hecho, sus consecuencias aún persisten. Algunos de los distritos escolares más importantes de Estados Unidos utilizaron el reporte como justificación para implementar ciertas políticas educativas controvertidas, tales como encomendar demasiadas ocupaciones a los niños.

La principal variable dependiente en este estudio era el rendimiento verbal. Hubo, sin embargo, más de 100 variables independientes. Los autores utilizaron procedimientos multivariados relativamente sofisticados para analizar los datos. Gran parte del foco de los problemas analíticos, la interpretación de los resultados y las críticas subsecuentes recaen en la naturaleza no experimental de la investigación.

La controvertida conclusión mencionada antes, sobre la importancia relativa de las variables de antecedentes del hogar y de las variables escolares, depende de un método completamente confiable y válido de evaluación de los impactos relativos de las diferentes variables. En la investigación experimental se está más a salvo al hacer conclusiones comparativas, ya que las variables independientes no están correlacionadas. Sin embargo, en el mundo educativo real las variables están correlacionadas, lo cual provoca que sea difícil determinar su contribución única al rendimiento. Aunque existen métodos estadísticos para manejar dichos problemas, ningún método puede decir, sin ambigüedades, que X_1 influye en Y en tal o cual medida porque la influencia real puede ser x_2, que influye tanto en x_1 como en y. La interpretación "correcta" de los hallazgos de *igualdad*, y de estudios similares, resulta siempre insostenible. Aunque existen métodos analíticos poderosos que se utilizan con datos no experimentales, las respuestas inequívocas a las preguntas sobre los factores determinantes del rendimiento son inalcanzables.

Diferencias del estilo de respuesta entre estudiantes del este asiático y estadounidenses

Esta investigación realizada por Chen, Lee y Stevenson (1995) fue un estudio de gran escala que abarcó cuatro países, cuatro culturas y dos continentes. El interés principal del

estudio se centró en el uso de escalas de medición, las cuales son básicas en la investigación de las ciencias del comportamiento. Sin embargo, ¿existen diferencias culturales respecto a la manera en que ciertos grupos étnicos responden las preguntas con el uso de una escala de medición? Chen y sus colaboradores deseaban determinar si existía o no una diferencia en el estilo de respuesta entre los asiáticos del este y los estadounidenses. Estos investigadores recolectaron datos de estudiantes: 944 japoneses, 1 357 taiwaneses, 687 canadienses y 2 174 del medio oeste y de la costa este de Estados Unidos. Las comparaciones en su estudio incluyeron las diferencias entre dos culturas (del este asiático contra estadounidense) y las diferencias entre los dos grupos representativos dentro de cada cultura (Estados Unidos contra Canadá, y Japón contra Taiwán). El cuestionario administrado a dichos estudiantes incluía reactivos sobre ideas, valores, actitudes, creencias y autoevaluaciones, en relación con la escuela y la vida cotidiana. Se utilizó una escala tipo Likert de 7 puntos, donde el 7 normalmente indicaba "más" o "fuertemente de acuerdo"; el 1 fue utilizado para "sin importancia", "menos" o "fuertemente en desacuerdo". Los resultados demostraron que en el área de individualismo y colectivismo hubo diferencias altamente significativas entre las dos culturas. También encontraron una relación entre el apoyo al individualismo y el uso de valores extremos en la escala.

Este estudio está clasificado también como no experimental ya que no hay una variable independiente manipulada. La variable independiente de este estudio era la cultura, y ésta no fue manipulada. El estudio mostró una diferencia en el estilo de respuesta entre diferentes culturas. Independientemente del muestreo de este estudio, éste es no experimental. Por lo tanto, no se puede afirmar de manera explícita que si usted es de esta cultura, entonces responderá de tal o cual forma en las escalas de medición.

Investigación no experimental a menor escala

Ilustrar estudios o series de estudios de investigación no experimental sobre el comportamiento no es fácil, pues existen demasiados; no obstante, algunos satisfacen los criterios personales de los autores sobre la solidez metodológica y el interés sustantivo. Se escogieron los siguientes estudios por tres razones: 1) los autores consideran que cada uno representa un enfoque único, original e interesante para un importante problema sociológico, psicológico o educativo; 2) cada uno de ellos contribuye significativamente al conocimiento científico, y 3) cada uno de ellos es no experimental.

Cochran y Mays: sexo, mentiras y VIH

Éste es el estudio clásico, frecuentemente citado y mencionado respecto a la diferencia en la conducta sexual entre hombres y mujeres. Cochran y Mays (1990) encontraron que aconsejar a los adultos jóvenes y a los adolescentes sobre las precauciones que deben tomar para protegerse del virus de inmunodeficiencia humana (VIH) puede resultar inútil. Por ejemplo, un consejo es que el individuo pregunte a la persona con quien va a salir, sobre su historia de riesgos, antes de decidir si se involucra o no en una relación sexual. Sin embargo, Cochran y Mays encontraron que la gente joven tiende a mentir respecto a su historia sexual. En una muestra de 665 estudiantes (entre 18 y 25 años de edad) en el sur de California, 196 hombres sexualmente experimentados y 226 mujeres sexualmente experimentadas reportaron que mentían para lograr relaciones sexuales, lo cual significa que más del 63% de la muestra afirmó haber mentido en el pasado, para tener sexo con quien salía. Se encontró que los hombres mentían significativamente más seguido que las mujeres. Además, tanto hombres como mujeres indicaron que volverían a engañar a su compañero o compañera de citas; los hombres estaban más dispuestos a hacer esto que las

TABLA 23.1 *Porcentaje de respuestas a preguntas sobre sexo y deshonestidad para hombres y mujeres (estudio de Cochran y Mays)*

Pregunta	Hombres	Mujeres
Ha mentido para tener relaciones sexuales	34	10
Mintió sobre el control en la eyaculación o la posibilidad de un embarazo	38	14
Mentiría respecto a tener resultados negativos de anticuerpos VIH	20	4
Mentiría sobre el control de la eyaculación o la posibilidad de un embarazo	29	2
Subestimaría el número de parejas previas	47	42

mujeres. En la tabla 23.1 se presenta un resumen del análisis realizado por Cochran y Mays sobre el cuestionario de 18 páginas respecto a la conducta sexual, la reducción del riesgo del VIH y del engaño en las citas. La comparación entre hombres y mujeres incluye la variable independiente medida o atributiva *sexo*. Dicho estudio posee gran significancia respecto a aconsejar a la gente joven sobre prácticas de sexo seguro. La implicación es que no se puede confiar en la palabra de la persona con quien se tiene una cita. Aunque los datos señalan fuertemente la disposición de los hombres a mentir para obtener favores sexuales, no es posible asumir automáticamente que todos los hombres mentirán sobre su historia sexual.

Elbert: problemas de lectura y del lenguaje escrito en niños con déficit de atención

El fenómeno conocido como trastorno por déficit de atención con hiperactividad (TDAH) es actualmente un área común de investigación psicológica y educativa. Los niños que la padecen por lo común exhiben escasa autorregulación de conductas y pobre desempeño escolar (generalmente de 1 a 1.5 unidades de desviación estándar por debajo de las puntuaciones de los niños normales). Conforme la investigación en esta área maduraba, los investigadores volcaron su interés a cuestiones más específicas que la comparación de los niños con TDAH con niños normales. Un área de interés son las subclases o subgrupos de TDAH, especialmente el trastorno por déficit de atención (TDA). Dichos estudios normalmente comparaban niños con déficit de atención e hiperactividad (TDA + H) contra niños con déficit de atención sin hiperactividad (TDA – H).

Uno de ellos fue realizado por Elbert (1993), quien deseaba determinar si estos dos subgrupos de TDAH (TDA + H y TDA – H) diferían respecto al rendimiento. El rendimiento se evaluaba a través de pruebas estandarizadas de lectura, ortografía y lenguaje escrito. Elbert también buscaba determinar si existía una interacción entre el género (hombre y mujer), la edad (6 a 7 años, 8 a 9 años y 10 a 12 años) y el tipo de subgrupo (TDA + H y TDA – H). El estudio utilizó los datos de 115 niños cuyas edades iban de los 6 a los 12 años. Cada niño era clasificado con TDA + H o TDA – H, por medio de evaluaciones objetivas de maestros y por los lineamientos establecidos por Barkley (1990). Note que aquí no hay una variable independiente manipulada. La inclusión en el grupo no se determinó a través de un proceso aleatorio. La naturaleza no experimental de tal estudio resultó en grupos de muy distintos tamaños. El grupo TDA + H quedó integrado por 83 niños; y el grupo TDA – H, por 32 niños. También había más hombres (86) que mujeres (29). Sin embargo, Elbert realizó numerosos análisis para verificar la igualdad de los grupos respecto a las variables edad, nivel de calificaciones, nivel de educación de la madre y CI. Las pruebas estadísticas entre TDA + H y TDA – H, con estas variables, no fueron significativas.

Los resultados en las puntuaciones de lectura mostraron un desempeño más pobre del grupo de niños con TDA + H (97), que en el grupo de TDA – H (97). Elbert también encontró una interacción significativa de género y edad, respecto a las puntuaciones de lectura. Pruebas *post hoc* demostraron que las niñas del grupo de edad media tuvieron un peor desempeño que los niños. Las pruebas estadísticas realizadas con las puntuaciones en cuanto a ortografía y lenguaje escrito no mostraron algún efecto entre los grupos de edad, género o tipo de subgrupo. No obstante, se encontró un efecto de interacción significativo del género y la edad. De nuevo, las niñas del grupo de edad media tuvieron el desempeño más pobre. Elbert descubrió además que una subprueba de ortografía y lenguaje escrito (ortografía escrita por dicción) fue la habilidad más limitada en ambos subgrupos de TDA. Observe la naturaleza no experimental del estudio de Elbert. No hubo variable independiente manipulada ni aleatorización.

Con estos estudios no experimentales de respaldo, ahora es posible discutir y evaluar la investigación no experimental, en general. Sin embargo, es necesaria una discusión evaluativa, con un cuestionamiento más sistemático de la comprobación de hipótesis alternativas, una de las características más importantes de la investigación científica.

Comprobación de hipótesis alternativas

La mayoría de las investigaciones se inician con las hipótesis; y después se prueban las implicaciones empíricas de tales hipótesis. Aunque las hipótesis se "confirman" de la manera descrita en capítulos anteriores, también se pueden "confirmar" y "desmentir" las hipótesis bajo estudio al tratar de demostrar qué hipótesis alternativas posibles son apoyadas o no. Primero, considere variables independientes alternativas como antecedentes de una variable dependiente. El razonamiento es el mismo. Por ejemplo, si se dice "variables independientes alternativas", en efecto se están proponiendo hipótesis o explicaciones alternativas de una variable dependiente.

En los estudios no experimentales, aunque no es posible confiar en la "verdad" de una proposición "si x, entonces y", como en un experimento verdadero, sí es posible establecer y probar hipótesis alternativas o "control". (Por supuesto, las hipótesis alternativas también pueden comprobarse, y de hecho así ocurre, en estudios experimentales.) Este procedimiento ha sido formalizado y explicado por Platt (1964), quien recibió la influencia de Chamberlin (1890; 1965). Platt le llama "inferencia fuerte". Chamberlin denomina apropiadamente al procedimiento "el método de trabajar hipótesis múltiples" y describe cómo pueden evitarse los propios "afectos intelectuales". Chamberlin (p. 756) asevera:

> El esfuerzo radica en sacar a la luz cada explicación racional de nuevos fenómenos e intentar desarrollar cada hipótesis sostenible, respetando su causa e historia. Así, el investigador se convierte en padre de una familia de hipótesis; y, por su relación parental con todas, el investigador tiene prohibido mostrar indebidamente su afecto a una de ellas.

Para revisar el desarrollo histórico de las hipótesis alternativas, véase Cowles (1989).

Sean x_1, x_2 y x_3 tres variables independientes alternativas, y sea y la variable dependiente, el fenómeno a "explicarse" con la proposición "si x, entonces y". Suponga que x_1, x_2 y x_3 agotan todas las posibilidades. Esta suposición no puede hacerse en realidad en la investigación científica, ya que resulta prácticamente imposible agotar todas las posibilidades. Aun así, aquí se supondrá esto por razones didácticas.

Un investigador tiene evidencia de que x_1 y y están sustancialmente relacionadas, y teniendo razones para creer que x_1 es el factor determinante, x_2 y x_3 se mantienen constantes. El supuesto aquí es que uno de los tres factores, x_1 o x_2 o x_3, es la "verdadera" variable independiente. De nuevo, note el supuesto. Puede ser ninguna o alguna combinación de

las tres. Suponga que el investigador tiene éxito en eliminar x_2; es decir, se demuestra que x_2 no está relacionada con y. Si el investigador también tiene éxito en eliminar x_3, entonces la conclusión es que x_1 es la variable independiente influyente. Puesto que las hipótesis alternativas o "control" no fueron justificadas, entonces se refuerza la hipótesis original.

De forma similar, es posible probar variables *dependientes* alternativas, lo que también implica hipótesis alternativas; se cambian las alternativas a la variable dependiente. Alper, Blane y Abrams (1955) lo ilustran en un estudio sobre las diferentes reacciones de niños de clase media y clase baja a pintar con los dedos, como consecuencia de diferentes prácticas de crianza del niño. La pregunta general planteada era: ¿las diferencias en la clase social en las prácticas de entrenamiento de niños resultarán en diferencias de personalidad entre clases? La teoría subyacente requería que hubiese diferencias en las reacciones ante el hecho de pintar con los dedos. Alper y sus colaboradores pensaron que los niños de clase media reaccionarían de manera distinta que los niños de clase baja a 16 variables diferentes, *cuando se utilizaran pinturas dactilares*: aceptación de la tarea, lavarse, etcétera. Las reacciones fueron significativamente diferentes respecto a la mayoría de las variables. En un "experimento control" se siguió el mismo procedimiento *utilizando crayolas en lugar de pinturas dactilares*. Los dos grupos no difirieron significativamente en ninguna de las 11 variables medidas. Éste fue un sorprendente contraste con los resultados de la pintura dactilar. El estudio fue no experimental ya que no era posible manipular la variable independiente, y porque los niños llegaron al estudio con sus reacciones ya construidas. El uso de un estudio control fue ingenioso y crucial. ¡Imagínese la consternación del investigador si las diferencias entre los dos grupos, en la tarea con crayolas, hubiesen resultado significativas!

Ahora considere el estudio clásico de Sarnoff, Lighthall, Waite, Davidson y Sarason (1958) que predijo que los niños ingleses y estadounidenses diferirían significativamente respecto a la ansiedad al contestar una prueba, pero no respecto a la ansiedad en general. La hipótesis fue delineada cuidadosamente: si se toma el examen eleven-plus, entonces resultará ansiedad de prueba. (El examen eleven-plus se aplica a los estudiantes ingleses a la edad de 11 años, como ayuda para determinar sus futuros educativos.) Puesto que era posible que hubiera otras variables independientes que causaran las diferencias entre los niños ingleses y estadounidenses, respecto a la ansiedad de prueba, evidentemente los investigadores quisieron excluir, por lo menos, a algunos de los principales competidores. Esto lo lograron apareando con cuidado las muestras: ellos quizás pensaron que la diferencia en la ansiedad de prueba podría deberse a una diferencia en la ansiedad en general, ya que la medida de la ansiedad de prueba obviamente debe reflejar alguna ansiedad en general. Si el resultado fuera éste, no se sustentaría la hipótesis principal. Por lo tanto, Sarnoff y sus colegas, además de comprobar la relación entre el examen y la ansiedad de prueba, también probaron la relación entre la examinación y la ansiedad general. La hipótesis de que los niños ingleses tendrían puntuaciones de ansiedad de prueba mayores que los estadunidenses fue confirmada por los datos. También encontraron que no hubo diferencias significativas entre los dos países respecto a la ansiedad en general, y que las niñas mostraban un nivel de ansiedad de prueba mayor que los niños, en ambos países. Se encontró que la ansiedad de prueba estaba correlacionada positivamente con el nivel de calificaciones.

Aunque el método de comprobación de hipótesis alternativas es importante en toda investigación, resulta fundamental en los estudios no experimentales, ya que es una de las únicas formas de controlar las variables independientes de dicha investigación. Al carecer de la posibilidad de la aleatorización y de la manipulación, los investigadores no experimentales, quizás más que los experimentales, deben ser muy sensibles a las posibilidades de prueba de hipótesis alternativas.

Evaluación de la investigación no experimental

El lector puede sentir, a causa de la discusión precedente, que la investigación no experimental es inferior que la experimental; pero esta conclusión sería injustificada. Es fácil afirmar que la investigación experimental es "mejor" que la investigación no experimental, o que la investigación experimental tiende a ser "trivial", o que la no experimental es "meramente correlacional". Dichas afirmaciones son, en sí mismas, simplificaciones excesivas. Lo que el estudiante de investigación necesita es una comprensión balanceada de las fortalezas y debilidades de ambos tipos de investigación. Estar comprometido inequívocamente con la investigación experimental o con la no experimental llega a convertirse en una visión limitada.

Limitaciones de la interpretación no experimental

La investigación no experimental posee tres grandes debilidades, dos de las cuales ya se analizaron en detalle: 1) la incapacidad de manipular variables independientes, 2) la falta de poder para aleatorizar y 3) el riesgo de realizar interpretaciones inadecuadas. En otras palabras, comparada con la investigación experimental, manteniendo lo demás igual, la investigación no experimental carece de control; tal carencia es la base de la tercera debilidad: el riesgo de la interpretación inadecuada.

El peligro de llegar a interpretaciones inadecuadas y erróneas en la investigación no experimental surge, en parte, de la posibilidad de muchas explicaciones para eventos complejos. Es fácil aceptar la primera y más obvia interpretación de una relación establecida, especialmente si se trabaja sin hipótesis que guíen la investigación. La investigación que no está guiada por hipótesis, o la investigación que se realiza "para averiguar cosas", con frecuencia es no experimental. La investigación experimental tiende más a basarse en hipótesis establecidas cuidadosamente.

Las hipótesis son predicciones de la forma si-entonces. En un experimento de investigación, la predicción se hace a partir de una x bien controlada, a una y. Si la predicción resulta verdadera, entonces se está relativamente a salvo al plantear la proposición condicional: "si x, entonces y". No obstante, en un estudio no experimental bajo las mismas condiciones, se está considerablemente menos a salvo al establecer la proposición condicional, por las razones discutidas anteriormente. Las protecciones cuidadosas son más importantes en el último caso, especialmente en la selección y comprobación de hipótesis alternativas, tales como la predicha carencia de relación entre el examen eleven-plus y la ansiedad en general, en el estudio de Sarnoff. Una relación predicha (o no predicha) en investigación no experimental puede resultar bastante espuria; aunque su plausibilidad y conformidad con la preconcepción puede volverla fácil de aceptar. Éste es un peligro en la investigación experimental, pero es menos peligroso que en la investigación no experimental, pues una situación experimental resulta mucho más fácil de controlar.

La investigación no experimental que se realiza sin hipótesis y sin predicciones, es decir, aquella en la cual los datos simplemente se recolectan y luego se interpretan, es aún más peligrosa por su capacidad de generar confusión. Si es posible, se localizan las diferencias o correlaciones significativas, y luego se interpretan. El segundo autor de este libro ha visto a estudiantes de posgrado recolectar una gran cantidad de datos sin hipótesis, y luego utilizar un programa de cómputo para realizar cualquier análisis posible, con la esperanza de encontrar significancia estadística en alguna parte. Después de encontrar diferencias significativas, se desarrollan las hipótesis que se adecuen al análisis. Para ilustrar el problema, suponga que un educador decide estudiar los factores que conducen a un bajo rendimiento académico. Se selecciona un grupo de sujetos con bajo rendimiento y un grupo de

personas con rendimiento normal. A cada grupo se le aplica una batería de pruebas. Después, se calculan las medias de las pruebas de los dos grupos, y las diferencias entre las medias se analizan mediante pruebas *t*. Digamos que de 12 diferencias, tres son significativas. Entonces, el investigador concluye que las personas con bajo rendimiento y las personas con rendimiento normal difieren respecto a las variables medidas con esas tres pruebas. Armado con estos análisis, el investigador ahora se siente deseoso de señalar a otros qué aspectos caracterizan a aquellos con bajo rendimiento. Puesto que las tres pruebas parecen medir inseguridad, entonces la causa del bajo rendimiento es la inseguridad.

Cuando están guiados por hipótesis, la credibilidad de los resultados de los estudios, como el descrito anteriormente, puede incrementarse; pero los resultados permanecen débiles porque se generan por el azar: debido únicamente al azar, uno o dos resultados de muchas pruebas estadísticas pueden ser significativos; sobre todo, la plausibilidad puede confundir. Una explicación plausible con frecuencia parece irresistible —¡aunque muy equivocada!—. Por ejemplo, parece muy obvio que los conservadores y los liberales son opositores; sin embargo, la evidencia de investigación parece indicar que no es así (Kerlinger, 1967, 1980, 1984). Otra dificultad es que las explicaciones plausibles, una vez halladas y creídas, a menudo son difíciles de probar. Según Merton (1949), las explicaciones *post factum* no conducen, por sí mismas, a la nulificación, ya que son demasiado flexibles. Cualesquiera que sean las observaciones, indica, pueden encontrarse nuevas interpretaciones que "se adecuen a los hechos" (pp. 90-91).

El valor de la investigación no experimental

A pesar de sus debilidades, en psicología, sociología y educación debe realizarse gran cantidad de investigación no experimental tan sólo porque muchos problemas de investigación no se prestan al cuestionamiento experimental. Una breve reflexión respecto a algunas de las variables importantes en investigación del comportamiento —inteligencia, aptitud, antecedentes familiares, rendimiento, clase social, rigidez, etnocentrismo— mostrará que no son manipulables. La búsqueda controlada es posible, por supuesto, aunque no la experimentación verdadera.

Incluso puede decirse que la investigación no experimental es más importante que la investigación experimental; ésta no es, por supuesto, una observación metodológica; más bien significa que la mayoría de los problemas de investigación científica social y educativa no conducen por sí mismos a la experimentación, aunque muchos de ellos conducen a la búsqueda controlada del tipo no experimental. Considere los estudios de Piaget sobre el pensamiento de los niños; los estudios de Adorno, Frenkel-Brunswik, Levinson y Sanford sobre el autoritarismo; el muy importante estudio de *Equality of Educational Opportunity*, y el estudio de Cochran y May sobre las mentiras y la práctica del sexo seguro. Considere además la influencia del estudio no experimental respecto del tabaquismo y de los problemas de salud: condujo a la creación de una legislación específica respecto de poner advertencias impresas en los propios productos. Si se hiciera un registro de estudios firmes e importantes en las ciencias del comportamiento y educación, es posible que los estudios no experimentales superaran en número y calidad a los estudios experimentales.

Conclusiones

Los estudiantes de investigación difieren ampliamente en sus puntos de vista respecto a los valores relativos de la investigación experimental y de la no experimental. Están aquellos

que exaltan la investigación experimental y subestiman la no experimental y aquellos que critican la discutida estrechez y la carencia de "realidad" de los experimentos, especialmente los de laboratorio. Dichas críticas, sobre todo en educación, enfatizan el valor y la relevancia de la investigación no experimental en situaciones "de la vida real" y "naturales" (como repaso consulte a Keith, 1988). Una posición racional parece obvia. Si es posible, utilice la experimentación porque, *si el resto de las cosas permanecen iguales*, es posible interpretar los resultados de la mayoría de los experimentos con gran confianza en que las proposiciones de la forma "si *p*, entonces *q*" son lo que se dice que son. También parece deseable probar las proposiciones "si *p*, entonces *q*" en otros escenarios. Se buscaría evidencia no experimental de la validez empírica de la propia hipótesis. Así, si es posible, las proposiciones condicionales deben estudiarse utilizando tanto el modelo experimental como el no experimental. Algunos estudios de investigación no experimental son impresionantes y convincentes. ¡Pero cuánto más impresionante y convincente sería si conclusiones similares surgieran de experimentos bien conducidos! De forma inversa, cuánto más convincentes son las conclusiones experimentales cuando se basan en investigación no experimental bien conducida.

La réplica es siempre deseable, incluso necesaria. Un aspecto importante a remarcar es que la réplica de la investigación no sólo significa repetición de los mismos estudios en las mismas situaciones. Podría y debería significar la comprobación de las implicaciones empíricas de la teoría —interpretando el término "teoría" en un sentido amplio— en situaciones similares y diferentes, y experimental y no experimentalmente. Es más fácil pedir extensiones de la investigación del laboratorio hacia el campo; pero los investigadores deben intentar concebir la comprobación experimental de proposiciones surgidas de manera no experimental. Por supuesto, esto es más difícil y rara vez se hace. Lo importante aquí es que *debe* concebirse y, cuando sea posible, hacerse.

El adoptar una posición firme de que la investigación experimental y no experimental es el único camino al cielo de la investigación es un asunto dogmático. Quizás sea muy difícil, incluso imposible en muchos casos, realizar tanto investigación experimental como no experimental en un mismo problema. ¿Será posible manipular experimentalmente la variable género de Cochran y May, o la variable cultural de Chen, Lee y Stevenson, por ejemplo? Por supuesto que difícil no quiere decir imposible. La cuestión aquí es que las posibilidades experimentales y las no experimentales deben ser exploradas y explotadas cuando sea posible hacerlo. Además, no debe asumirse de manera inmediata que no es posible realizar investigación de manera distinta a como se ha hecho. No existe un solo camino metodológico hacia la validez científica; existen muchos. Los caminos deben elegirse por su adecuación a los problemas bajo estudio. Sin embargo, esto no quiere decir que no se pueda explotar un modelo que difiere de aquello a lo que se está acostumbrado.

Por alguna extraña razón, quizás la creencia espuria en la supuesta certeza de la ciencia, cuando la gente —incluyendo a los científicos— piensa en la ciencia y en la investigación científica, se considera erróneamente que tan sólo existe una forma "correcta" de acercamiento y de hacer investigación. Muy raras veces se comete dicho error en la música, el arte o en la construcción de una casa. También la ciencia tiene muchos caminos, y los modelos experimental y no experimental son dos de ellos. Ninguno de ellos es correcto o erróneo; sino que son diferentes. La tarea aquí ha sido tratar de entender las diferencias y sus consecuencias. Sin embargo, todavía falta mucho para finalizar el tema. Es probable que se logre más comprensión antes de finalizar. Cuando se piense en los diferentes puntos de vista de los métodos experimentales y los no experimentales, debe considerarse la máxima china que afirma que "existen muchos caminos hacia la cima de la montaña, pero la vista ahí siempre es la misma".

RESUMEN DEL CAPÍTULO

1. Cuando se conducen de forma correcta los estudios no experimentales son tan valiosos como los experimentales.
2. Un ingrediente para un buen estudio no experimental es el desarrollo de hipótesis antes de iniciar el estudio.
3. La réplica sirve para incrementar la credibilidad de los resultados obtenidos a partir de estudios no experimentales.
4. La investigación no experimental se define como aquella que no posee una variable independiente activa.
5. La diferencia más importante entre los métodos experimental y no experimental es el control.
6. La autoselección de los participantes es un problema importante de los estudios no experimentales.
7. Existe un gran número de estudios no experimentales realizados y publicados en las ciencias del comportamiento.
8. Algunos estudios no experimentales —como el que relacionó el tabaquismo con los problemas de salud— han tenido una gran influencia.
9. Existen tres debilidades principales en la investigación no experimental: (i) las variables independientes no pueden ser manipuladas, (ii) la carencia de aleatorización, y (iii) el riesgo de la interpretación inadecuada.
10. La eliminación paso a paso de las hipótesis alternativas es una forma de llegar a la variable que posiblemente "causa" los cambios en la variable dependiente.
11. Desarrollos relativamente nuevos en estudios no experimentales incluyen el modelo matemático de "causa y efecto". Estos modelos, en realidad, no implican causa y efecto.

SUGERENCIAS DE ESTUDIO

1. Un psicólogo social planea investigar los factores que subyacen al antisemitismo. Él considera que las personas que tuvieron padres autoritarios y una educación autoritaria tienden a ser antisemitas. ¿Un proyecto de investigación diseñado para probar esta hipótesis sería experimental o no experimental? Explique.
2. Un psicólogo educativo decide probar la hipótesis de que la inteligencia y la motivación son los principales factores determinantes del éxito escolar. ¿Esta investigación sería principalmente experimental o no experimental? Argumente.
3. Un investigador está interesado en la relación entre la percepción del papel y los valores sociales.
   a) ¿Cuál es la variable independiente? ¿Y la variable dependiente?
   b) Cualquiera que haya sido su juicio, ¿puede invertir de forma justificada las variables?
   c) ¿Cree usted que un proyecto de investigación diseñado para investigar este problema sería básicamente experimental o no experimental?
   d) ¿Puede el investigador hacer dos investigaciones, una experimental y otra no experimental, ambas diseñadas para probar la misma hipótesis?
   e) Si su respuesta para *d)* fue "Sí", ¿serían las mismas variables en los dos problemas? Suponiendo que las relaciones en ambas investigaciones fuesen significativas, ¿las conclusiones serían sustancialmente las mismas?

TABLA 23.2 *Países con alta y baja motivación de logro, cuya producción de energía eléctrica estuvo por arriba o por debajo de las expectativas (estudio de McClelland)*[a]

	Por arriba de las expectativas	Por debajo de las expectativas
Alta motivación de logro	13 (65%)	7 (35%)
Baja motivación de logro	5 (26%)	14 (70%)

[a] Los datos en las casillas son el número de países que, por ejemplo, tuvieron alta motivación de logro y cuya producción de electricidad estuvo por arriba de las expectativas (13). Los índices en los paréntesis son los porcentajes.

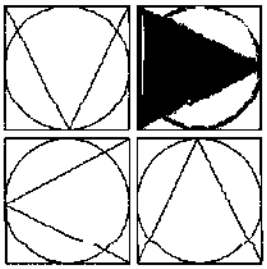
¿Apoyan estos resultados la hipótesis de McClelland? (Sugerencia: Calcule χ^2 y C, como en el capítulo 10. Utilice los porcentajes para ayudarse a interpretar la tabla.)

[Respuesta: $\chi^2 = 5.87$, $gl = 1$ ($p < .05$); C = .36. Sí, la hipótesis se mantiene.]

4. En las sugerencias de estudio del capítulo 2 se presentaron una serie de problemas e hipótesis. Tome cada uno de estos problemas e hipótesis y decida si la investigación diseñada para explorar los problemas y para probar las hipótesis serían básicamente experimentales o no experimentales. ¿Puede alguno de los problemas e hipótesis tratarse de ambas maneras?
5. McClelland (1961) presenta datos sobre la producción de energía eléctrica durante los años 1952-1958 de países con alta motivación de logro y baja motivación de logro. Al contar el número de países en cada una de las cuatro casillas se obtienen los datos que se muestran en la tabla 23.2.
6. El estudiante emprendedor quizás desee dar un paso decisivo hacia un pensamiento estimulante, provocativo, controvertido e importante. El famoso reporte del Club de Roma de Meadows, Meadows, Randers y Behrens (1974) molestó a algunos observadores, alarmó casi a todos los que lo han leído y ofendió a todos. Con el uso de variables sociales importantes —recursos naturales, contaminación, población, por ejemplo— y sus interacciones complejas, se ha pronosticado un desastre final para las ciudades y el mundo. La investigación en que se basan las conclusiones es totalmente no experimental. Lea este reporte. ¿Considera que el carácter no experimental de la investigación disminuye su credibilidad?
7. Lea uno de (o todos) los siguientes estudios. Todos son no experimentales. Escriba las razones por las que usted piensa que son no experimentales con base en los puntos resaltados en este capítulo.

Goodman, S. H. y Emory, E. K. (1992). Perinatal complications in births to low socioeconomic status schizophrenic and depressed women. *Journal of Abnormal Psychology, 101*, 225-229.

Koniak-Griffin, D. y Brecht, M. (1995). Linkages between sexual risk taking, substance use, and AIDS knowledge among pregnant adolescents and young mothers. *Nursing Research, 44*, 340-346.



Capítulo 24

Experimentos de laboratorio, experimentos de campo y estudios de campo

- Experimento de laboratorio: estudios de Miller del aprendizaje de respuestas viscerales
- Un experimento de campo: el estudio de Rind y Bordia sobre los efectos del agradecimiento de un mesero y la personalización en las propinas de los restaurantes
- Un estudio de campo: el estudio de Bennington College realizado por Newcomb
  - Características y criterios de los experimentos de laboratorio, experimentos de campo y estudios de campo
  - Fortalezas y debilidades de los experimentos de laboratorio
  - Propósitos de los experimentos de laboratorio
  - El experimento de campo
  - Fortalezas y debilidades de los experimentos de campo
  - Estudios de campo
  - Tipos de estudios de campo
  - Fortalezas y debilidades de los estudios de campo
- Investigación cualitativa
- Anexo: el paradigma experimental holístico

La investigación científica social puede dividirse en cuatro grandes categorías: experimentos de laboratorio, experimentos de campo, estudios de campo e investigación mediante encuestas. Esta clasificación surge de dos fuentes: la distinción entre investigación experimental y no experimental, y la distinción entre investigación de laboratorio y de campo. Este capítulo le debe mucho al material de tres libros: 1) Festinger y Katz (1953),

2) Taylor y Bodgan (1998) y 3) Padgett (1998). Aunque la publicación de Festinger y Katz tiene más de 45 años, continúa siendo una fuente valiosa sobre muchos aspectos de la metodología de investigación del comportamiento. Este capítulo inicia con la presentación de ejemplos del experimento de laboratorio, dos experimentos de campo y un estudio de campo. Esto es para que el lector perciba los principales componentes de cada método y las diferencias entre cada uno.

Experimento de laboratorio: estudios de Miller del aprendizaje de respuestas viscerales

Una serie de brillantes experimentos realizados por Miller (1969, 1971) ha trastornado una creencia mantenida por largo tiempo: que el aprendizaje ocurre tan sólo con respuestas voluntarias, y que el sistema autónomo involuntario está sujeto únicamente al condicionamiento clásico. Esto, en efecto, indica que respuestas como mover la mano y hablar pueden controlarse y, por lo tanto, enseñarse; pero que las respuestas involuntarias, como la frecuencia cardiaca, las contracciones intestinales y la presión sanguínea no pueden tenerse bajo control instrumental y, por lo tanto, no pueden "enseñarse". Para entender los estudios de Miller es necesario definir ciertos términos psicológicos. En el *condicionamiento clásico* un estímulo neutral, inherentemente incapaz de producir cierta respuesta, se vuelve capaz de hacerlo al ser asociado repetidamente con un estímulo inherentemente capaz de producirla. El ejemplo más famoso es el del perro de Pavlov que salivaba ante el sonido de un metrónomo, que había sido asociado repetidamente con carne en polvo. En el *condicionamiento instrumental* u *operante*, dar un reforzamiento a un organismo, inmediatamente después de haber dado una respuesta, produce un incremento de la respuesta. Recompense una respuesta y ésta se repetirá. Se enseña que las respuestas o conductas voluntarias son superiores, quizás porque están bajo el control del individuo; mientras que las respuestas involuntarias son inferiores porque no son controladas. Se ha considerado que las respuestas involuntarias pueden modificarse únicamente por medio del condicionamiento clásico y no por medio del condicionamiento instrumental. En otras palabras, la posibilidad de "enseñar" al corazón, estómago o a la sangre es remota, ya que las situaciones del condicionamiento clásico son difíciles de lograr. No obstante, si se sujeta a los órganos al condicionamiento instrumental, pueden ponerse bajo control experimental y pueden ser "enseñados"; y ellos pueden "aprender".

El trabajo de Miller ha demostrado que, a través de condicionamiento instrumental, la frecuencia cardiaca es susceptible de modificarse, las contracciones estomacales llegan a alterarse e inclusive la formación de orina ¡puede incrementarse o disminuirse! Tal descubrimiento tiene una importancia teórica y práctica enorme. Para mostrar la naturaleza de los experimentos de laboratorio se tomará uno de los experimentos más interesantes y creativos de Miller.

La idea del experimento resulta simple: recompensar a un grupo de ratas cuando su frecuencia cardiaca aumente, y recompensar a otro grupo cuando su frecuencia cardiaca disminuya. Éste es un ejemplo directo del diseño de dos grupos analizado anteriormente. El gran problema de Miller era el control. Existen muchas otras causas del cambio en la frecuencia cardiaca, por ejemplo, el esfuerzo muscular. Para controlar dichas variables extrañas, Miller y su colega (Trowill) paralizaron a las ratas con curare. Pero si las ratas estaban paralizadas, ¿qué podría utilizarse como recompensa? Decidieron utilizar estimulación eléctrica directa en el cerebro. La variable dependiente, frecuencia cardiaca, se registraba de manera continua con el electrocardiógrafo. Cuando ocurría un pequeño

cambio en la frecuencia cardiaca (en la dirección "correcta": incremento para un grupo y decremento para el otro), se le aplicaba un impulso eléctrico al animal en un centro de recompensa de su cerebro (véase también Olds y Fobes, 1981, una investigación cerebral que demuestra que una pequeña estimulación eléctrica de cierta parte del cerebro actúa como recompensa). Esto se continuó hasta que los animales estuvieron "entrenados".

Los incrementos y decrementos de la frecuencia cardiaca fueron confiables estadísticamente, pero pequeños: sólo 5% en cada dirección. Por ello, Miller y otro colega (DiCara) utilizaron la técnica conocida como moldeamiento que, en este caso, implicó recompensar primero pequeños cambios y después exigir cambios crecientes en la frecuencia para obtener las recompensas. Esto incrementó los cambios en la frecuencia cardiaca a un promedio de 20% en cada dirección. Además, investigación posterior, donde se utilizó el escape de un choque eléctrico leve como reforzamiento, demostró que los animales recordaron lo que habían aprendido y que "diferenciaron" las respuestas cardiacas de otras respuestas.

Miller ha tenido éxito al "entrenar" otras respuestas involuntarias: contracción intestinal, formación de orina y presión sanguínea, por ejemplo. En resumen, las respuestas viscerales *pueden aprenderse* y *pueden moldearse*; ¿pero puede utilizarse este método con personas? Miller dice que él considera que las personas son tan inteligentes como las ratas, pero que no ha sido completamente probado todavía. Aunque el uso de curare podría presentar dificultades, Miller dice que puede hipnotizarse a la gente.

Un experimento de campo: el estudio de Rind y Bordia sobre los efectos del agradecimiento de un mesero y la personalización en las propinas de los restaurantes

¿La práctica común entre los meseros de escribir "gracias" al reverso de la cuenta y entregarla, de tal manera que el comensal vea la gratitud del mesero, produce mayores propinas? Si en realidad es así, entonces el mesero se beneficia con esta acción a un costo extremadamente bajo. Rind y Bordia (1995) realizaron este experimento de campo para determinar la efectividad de utilizar dicha técnica: específicamente, escribir "gracias" y personalizar la interacción mesero-comensal añadiendo el nombre del mesero. El estudio fue realizado en un elegante restaurante de Filadelfia durante el periodo del almuerzo, por cinco días. Participaron 51 comensales en el estudio. Todas las meseras eran del sexo femenino. La variable independiente era la *impresión* y consistió de tres niveles: 1) la parte posterior de la cuenta no contenía nada, 2) la palabra "gracias" escrita a mano, o 3) la palabra "gracias" más el nombre de pila de la mesera. Rind y Bordia hipotetizaron, con base en la teoría del manejo de la impresión, que el añadir el agradecimiento escrito y personalizado conduciría a propinas más altas, que si la mesera no escribía nada al reverso de la cuenta. También hipotetizaron que la personalización de la cuenta conduciría a propinas más altas que sin la personalización. Cada nivel o condición de la variable independiente se determinó aleatoriamente para cada grupo de comensales. Antes de entregar la cuenta al comensal, la mesera tomaba aleatoriamente de su bolsillo una de tres monedas (fechadas 1981, 1982 y 1983). Si tomaba la moneda de 1981, la mesera no escribía nada al reverso de la cuenta. Si seleccionaba la moneda de 1982, anotaba "gracias" al reverso de la cuenta. Si la moneda elegida era la de 1983, escribía "gracias" al reverso de la cuenta, pero también añadía su nombre. Se registró el tamaño de la propina, el tamaño de la cuenta, el tamaño del grupo de comensales y la forma de pago, para cada grupo de comensales. Los resultados de dicho estudio demostraron que agregar la palabra "gracias"

a la cuenta producía propinas significativamente mayores que cuando no se escribía nada en la cuenta (18% de las cuentas contra 16.3%). No hubo diferencias significativas entre el agradecimiento escrito y el agradecimiento escrito más la personalización. Rind y Bordia mencionan que existen teorías encontradas respecto al porqué de sus hallazgos. Sin embargo, a partir de sus resultados, parece que esta práctica es benéfica para el mesero. Los investigadores también señalan las limitaciones de su experimento. Una de ellas es que su elección de conducir el estudio en un restaurante elegante puede generar resultados diferentes que si el estudio se hubiese realizado en un restaurante popular. El uso de mujeres únicamente abre la posibilidad de que los clientes pudieran tratar a los meseros hombres de manera diferente.

Un estudio de campo: el estudio de Bennington College realizado por Newcomb

Newcomb (1943) llevó a cabo uno de los estudios clásicos más importantes respecto a la influencia del ambiente universitario sobre los estudiantes. En él examinó al cuerpo entero de estudiantes del Bennington College (cerca de 600 mujeres jóvenes), de 1935 a 1939. Una faceta inusual del estudio fue el intento de Newcomb de explicar la influencia tanto de los factores personales como de los sociales sobre los cambios de actitud de los estudiantes. Aunque también se probaron otras hipótesis, la hipótesis principal del estudio de Bennington fue que los nuevos estudiantes coincidirían respecto a las normas del grupo universitario, y que cuanto más se integraran a la comunidad universitaria, mayor sería el cambio en sus actitudes sociales.

Newcomb utilizó varias escalas de actitud de papel y lápiz, informes escritos sobre los estudiantes y entrevistas individuales. El estudio fue longitudinal y no experimental. La variable independiente, que aunque no fue fácil de categorizar, puede decirse que fueron las normas sociales del Bennington College. La variable dependiente fueron las actitudes sociales y ciertos comportamientos de las estudiantes.

Newcomb encontró cambios significativos en las actitudes entre las estudiantes de nuevo ingreso, por un lado, y las estudiantes de niveles intermedios y de último año, por el otro. Los cambios se dieron hacia un menor conservadurismo en una variedad de aspectos sociales. Por ejemplo, las preferencias políticas de los estudiantes intermedios y del último año, en la elección presidencial de 1936, fueron mucho menos conservadoras que las de hombres jóvenes y estudiantes de segundo año. De 50 estudiantes intermedios y del último año, 15% preferían a Landon (republicano), mientras que de 52 estudiantes de nuevo ingreso, 62% preferían a Landon. Los porcentajes de preferencia por Roosevelt (demócrata) fueron 54% y 29%. Las puntuaciones medias de todas las estudiantes de los cuatro años, en una escala diseñada para medir conservadurismo político y económico, fueron: estudiantes de nuevo ingreso, 74.2; estudiantes de segundo año, 69.4; intermedios, 65.9, y estudiantes del último año, 62.4. Evidentemente la universidad afectó las actitudes de las estudiantes.

Newcomb planteó una pregunta de "control": ¿habrían cambiado estas actitudes en otras universidades? Para responder esta pregunta, administró sus medidas de conservadurismo a estudiantes del Williams College y del Skidmore College. Las puntuaciones medias comparativas de los estudiantes de Skidmore —desde los estudiantes de recién ingreso hasta los de último año— fueron 79.9, 78.1, 77.0 y 74.1. Parece que los estudiantes de Skidmore (y Williams) no cambiaron tanto, ni tan consistentemente a través del tiempo, como lo hicieron los de Bennington.

Newcomb, Koenig, Flacks y Warwick (1967) reportaron un estudio de seguimiento de los estudiantes del Bennington College, después de 25 años. Encontraron que los cambios habían perdurado y que la influencia de Bennington era persistente.

Características y criterios de los experimentos de laboratorio, experimentos de campo y estudios de campo

Un *experimento de laboratorio* es una investigación en la que la varianza de todas, o de casi todas, las posibles variables independientes influyentes, sin pertinencia al problema de investigación inmediato, se mantienen al mínimo. Esto se logra aislando la investigación en una situación física separada de la rutina de la vida ordinaria, y por medio de manipular una o más variables independientes bajo condiciones rigurosamente específicas, operacionalizadas y controladas.

Fortalezas y debilidades de los experimentos de laboratorio

El experimento de laboratorio tiene la virtud inherente de la posibilidad de un control relativamente completo. El experimento de laboratorio puede, y frecuentemente lo hace, aislar la situación de investigación de la vida fuera del laboratorio, al eliminar las muchas influencias extrañas que lleguen a afectar las variables independiente y dependiente.

Además del control de la situación, los experimentos de laboratorio generalmente se sirven de la asignación aleatoria y manipulan una o más variables independientes. Existen otros aspectos del control de laboratorio: en la mayoría de los casos, el investigador logra un alto grado de especificidad en las definiciones operacionales de las variables. Las relativamente crudas definiciones operacionales de las situaciones de campo, tales como muchas de las que se asocian con la medición de valores, actitudes, aptitudes y características de personalidad, no saturan al experimentador; aunque el problema de definición nunca resulta simple. El experimento de Miller (1969, 1971) constituye un buen ejemplo. Las definiciones operacionales del reforzamiento y del cambio en la frecuencia cardiaca son precisas y altamente objetivas.

Muy relacionada con la fortaleza operacional está la precisión de los experimentos de laboratorio. Preciso significa exacto, definido y no ambiguo. Las mediciones precisas se efectúan con instrumentos de precisión. En términos de varianza, cuanto más preciso sea un procedimiento experimental, menor será la varianza del error. Cuanto más exacto o preciso sea un instrumento de medición, mayor certeza se tendrá de que las medidas obtenidas no varían mucho de sus valores "verdaderos".

Los resultados precisos de laboratorio se logran principalmente por medio de la manipulación y medición controladas, en un ambiente donde se eliminaron las posibles condiciones "contaminantes". Los reportes de investigación de experimentos de laboratorio generalmente especifican en detalle la manera en que se realizaron las manipulaciones y los medios utilizados para controlar las condiciones del ambiente bajo las cuales se efectuaron. Al especificar de manera exacta las condiciones del experimento, se reduce el riesgo de que los participantes puedan responder erróneamente y, por lo tanto, de introducir varianza aleatoria a la situación experimental. El experimento de Miller constituye un modelo de precisión experimental de laboratorio.

La mayor debilidad del experimento de laboratorio probablemente es la carencia de fortaleza de las variables independientes. Puesto que las situaciones de laboratorio son, después de todo, situaciones creadas para propósitos especiales, puede decirse que los

efectos de las manipulaciones experimentales son generalmente débiles. Los incrementos y los decrementos en la frecuencia cardiaca por medio del reforzamiento eléctrico del cerebro fueron, aunque asombrosas, relativamente pequeñas. Compare esto con los efectos relativamente grandes de las variables independientes en situaciones reales. En el estudio Bennington, por ejemplo, la comunidad universitaria aparentemente tuvo un efecto masivo. En la investigación de laboratorio sobre la conformidad, generalmente se producen tan sólo pequeños efectos por la presión grupal sobre los individuos. Confronte lo anterior con el efecto relativamente fuerte de la mayoría de un gran grupo sobre un miembro individual de un grupo, en una situación de la vida real. El miembro del consejo de educación que sabe que cierta acción deseable va en contra de los deseos de la mayoría de sus colegas, y quizás de la mayoría de la comunidad, se encuentra bajo mucha presión para coincidir con la norma.

Una razón de la preocupación respecto a la precisión del laboratorio y de la estadística refinada es la debilidad de los efectos de laboratorio. Detectar una diferencia significativa en el laboratorio requiere de situaciones y medidas con el menor ruido debido al azar, y de pruebas estadísticas precisas y sensibles que muestren relaciones y diferencias significativas cuando existan.

Otra debilidad es un producto de la primera: la artificialidad de la situación de investigación experimental. En realidad, es difícil saber si la artificialidad es una debilidad o simplemente una característica neutral de las situaciones experimentales de laboratorio. Cuando una situación de investigación se idea deliberadamente para excluir las muchas distracciones del ambiente, quizá resulte ilógico etiquetar a la situación con un término que exprese en parte el resultado que se busca. La crítica sobre la artificialidad no proviene de los experimentadores que saben que las situaciones experimentales son artificiales; sino proviene de los individuos que no comprenden los propósitos de los experimentos de laboratorio.

La tentación de interpretar de forma incorrecta los resultados de los experimentos de laboratorio es enorme. Mientras que los resultados de Miller son considerados altamente significativos por los científicos sociales, sólo pueden ser extrapolados de manera tentativa a situaciones fuera del laboratorio. Resultados similares pueden obtenerse en situaciones de la vida real, y existe evidencia de que así sucede en algunos casos; pero esto no es así necesariamente. La relación debe siempre probarse una y otra vez bajo situaciones que no sean de laboratorio. Por ejemplo, la investigación de Miller tendrá que realizarse cuidadosa y precavidamente con seres humanos en hospitales e inclusive en escuelas.

A pesar de que los experimentos de laboratorio tienen una validez interna relativamente alta, carecen de validez externa. Antes se planteó la pregunta: ¿hizo realmente *X*, la manipulación experimental, una diferencia significativa? A mayor confianza en la "verdad" de las relaciones descubiertas en un estudio de investigación, mayor será la validez interna del estudio. Cuando se descubre una relación en un experimento de laboratorio bien realizado, por lo común se tiene bastante confianza en él, pues se ha ejercido el máximo control posible de la variable independiente y de otras posibles variables independientes extrañas. Cuando Miller "descubrió" que las respuestas viscerales podían aprenderse y moldearse, quizás estuvo relativamente seguro de la "verdad" de la relación entre el reforzamiento y la respuesta visceral en el laboratorio. Él consiguió un alto grado de control y de validez interna.

Puede decirse que si se estudia un problema utilizando experimentos de campo, *quizá* se encontrará alguna relación. Ésta es una cuestión empírica, no especulativa; la relación que se desea probar debe ponerse en la situación donde se quiere generalizar. Si un investigador encuentra que los individuos coinciden respecto a las normas grupales en el laboratorio, ¿ocurrirá el mismo fenómeno o uno similar en los grupos comunitarios, facultades

y cuerpos legislativos? Esta carencia de validez externa forma la base de las objeciones de muchos educadores respecto a los estudios con animales de las teorías de aprendizaje. Sus objeciones son válidas únicamente si un experimentador generaliza, a partir del comportamiento y aprendizaje de animales de laboratorio, al comportamiento y aprendizaje de los niños. Sin embargo, los experimentalistas capaces rara vez cometen un error de este tipo —ellos saben que el laboratorio constituye un ambiente restringido—.

Propósitos del experimento de laboratorio

Los experimentos de laboratorio tienen tres propósitos relacionados. Primero, son un medio para estudiar las relaciones bajo condiciones "puras" y no contaminadas. Los experimentadores se plantean las siguientes preguntas: ¿está x relacionada con y? ¿Cómo se relaciona con y? ¿Qué tan fuerte es la relación? ¿Bajo qué circunstancias cambia la relación? Ellos buscan escribir ecuaciones de la forma $y = f(x)$, hacer predicciones con base en la función, y ver qué tan bien y bajo qué condiciones se lleva a cabo la función.

Un segundo propósito debe mencionarse en conjunción con el primer propósito: en un inicio hay que comprobar las predicciones derivadas de la teoría y, después, aquellas derivadas de otras investigaciones.

Un tercer propósito de los experimentos de laboratorio consiste en refinar las teorías y la hipótesis, para formular hipótesis relacionadas con otras hipótesis probadas experimental o no experimentalmente y, quizás lo más importante, ayudar a la construcción de sistemas teóricos. Éste fue uno de los principales propósitos de Miller. Aunque algunos experimentos de laboratorio se realizan sin dicho propósito, la mayoría de los experimentos de laboratorio están, por supuesto, orientados a la teoría.

Entonces, el propósito de los experimentos de laboratorio consiste en probar hipótesis derivadas de la teoría, estudiar las interrelaciones precisas de variables y su operación, y controlar la varianza bajo condiciones de investigación que no estén contaminadas por la operación de variables extrañas. Como tal, el experimento de laboratorio es uno de los más grandes inventos de todos los tiempos. A pesar de que existen debilidades, lo son sólo en un sentido realmente irrelevante. Admitiendo la carencia de representatividad (validez externa), el experimento de laboratorio bien realizado aún cumple el prerrequisito fundamental de cualquier investigación: la validez interna.

El experimento de campo

Un experimento de campo consiste en un estudio de investigación realizado en una situación real, donde una o más variables independientes son manipuladas por el experimentador bajo condiciones tan cuidadosamente controladas como la situación lo permita. El contraste entre el experimento de laboratorio y el experimento de campo no es grande: las diferencias son principalmente cuestiones de grado. Algunas veces resulta difícil etiquetar un estudio particular como "experimento de laboratorio" o como "experimento de campo". En tanto que el experimento de laboratorio tiene un control máximo, la mayoría de los experimentos de campo deben operar con menos control, un factor que a menudo constituye una severa limitante.

Fortalezas y debilidades de los experimentos de campo

Los experimentos de campo poseen valores que los recomiendan especialmente para los psicólogos sociales, sociólogos y educadores, a causa de que se ajustan de forma admirable

a muchos de los problemas sociales y educativos de interés para la psicología social, la sociología y la educación. Puesto que las variables independientes se manipulan y se utiliza la aleatorización, puede cumplirse el criterio de control —por lo menos teóricamente—.

No obstante, el control de la situación experimental de campo rara vez resulta tan rígido como el del laboratorio. Esto implica tanto una fortaleza como una debilidad. En un experimento de campo, aunque el investigador tiene el poder de la manipulación, siempre se enfrenta con la desagradable posibilidad de que las variables independientes estén contaminadas por variables ambientales no controladas. Se enfatiza este punto pues la necesidad de controlar variables independientes extrañas es particularmente crítica en los experimentos de campo. El experimento de laboratorio se lleva a cabo en una situación altamente controlada; mientras que el experimento de campo se realiza en una situación natural y frecuentemente laxa. Por lo tanto, una de las preocupaciones principales del experimentador de campo consiste en tratar de hacer que la situación de investigación se aproxime lo más posible a las condiciones del experimento de laboratorio. Por supuesto que con frecuencia ésta es una meta difícil de lograr, pero si la situación de investigación puede mantenerse rígida, entonces el experimento de campo es poderoso porque, en general, es posible tener mayor confianza en que las relaciones en realidad sean lo que se dice que son.

Como compensación para la disminución del control, el experimento de campo posee dos o tres virtudes únicas. Las variables de un experimento de campo por lo común tienen un efecto más poderoso que las de los experimentos de laboratorio. Los efectos de los experimentos de campo a menudo son lo suficientemente fuertes para penetrar las distracciones de las situaciones experimentales. El principio es: *cuanto más realista sea la situación de investigación, más fuertes serán las variables.* Ésta es una de las ventajas de realizar investigación en ambientes educativos. En su mayoría, la investigación en los ambientes educativos es similar a las actividades educativas y, por lo tanto, no debe verse necesariamente como algo especial y separado de la vida escolar. A pesar de la petición de muchos educadores porque se realice investigación educativa más realista, no existe virtud especial en el realismo por sí mismo. El realismo tan sólo incrementa la fortaleza de las variables; también contribuye a la validez externa, ya que a mayor realismo de la situación, habrá mayor posibilidad de que las generalizaciones a otras situaciones resulten más válidas.

Otra virtud de los experimentos de campo es que son apropiados para estudiar influencias, procesos y cambios sociales y psicológicos complejos, en situaciones similares a la vida real. Glick, DeMorest y Hotze (1988), por ejemplo, estudiaron los efectos de la pertenecia al grupo, del espacio personal y de las solicitudes de ayuda sobre la ansiedad interpersonal y la obediencia de los individuos. Los investigadores realizaron su estudio en una plaza comercial, utilizando compradores reales como participantes. Schmitt, Dube y Leclerc (1992) estudiaron un problema similar sobre el espacio personal, examinando intrusiones en filas de espera. Estos investigadores condujeron tres experimentos de laboratorio y uno de campo, como un intento para determinar si las reacciones conductuales a la intrusión se basan en intereses personales o sociales. Jaffe (1991) realizó un experimento de campo sobre anuncios comerciales dirigidos a las mujeres. En dicha investigación los participantes evaluaron un anuncio impreso que incluía mujeres en diferentes situaciones. La situación utilizada era la de la mujer tradicional (nutriente y orientada a la familia) o la mujer moderna (exitosa en su carrera y en la familia). Rabinowitz, Colmar, Elgie, Hale, Niss, Sharp y Sinclito (1993) estudiaron la conducta compleja de los cajeros en tiendas de recuerdos de viaje que atienden a turistas. Los investigadores deseaban saber si el mal manejo del dinero se debía a la deshonestidad, la indiferencia o al descuido. El estudio de Wogalter y Young (1991) sobre la eficacia de advertencias verbales o impresas en el manejo de sustancias peligrosas, o en el piso resbaloso de una plaza comercial, son útiles para

quienes se preocupan de los aspectos de seguridad en ambientes industriales o comerciales. Wogalter y Young hicieron dos estudios de laboratorio y un experimento de campo para demostrar que la combinación de advertencias verbales e impresas era lo más eficaz para producir la conducta de obediencia en la gente. Todos estos estudios utilizaron manipulación experimental en los participantes de la vida real, en ambientes reales.

Los experimentos de laboratorio son adecuados principalmente para comprobar aspectos de teorías; mientras que los experimentos de campo son adecuados tanto para comprobar hipótesis derivadas de teorías, como para encontrar respuestas a problemas prácticos. Los experimentos sobre métodos educativos, generalmente con propósitos prácticos, con frecuencia buscan cuál de dos o tres métodos es el mejor para lograr cierto propósito. La investigación industrial y la investigación del consumidor dependen en gran parte de los experimentos de campo. Por otro lado, gran parte de la investigación en psicología social es básicamente teórica. El estudio de Schmitt y sus colaboradores (1992), mencionado anteriormente, probó dos teorías sobre las reacciones conductuales de las personas que, formadas en filas, experimentan una intrusión. Las dos teorías probadas fueron la teoría del ultraje moral y la teoría del costo individual. Los experimentos de campo de Glick y colaboradores (1988) y de Jaffe (1991), también estuvieron orientados a la teoría.

La flexibilidad y la aplicabilidad a una gran variedad de problemas constituyen rasgos importantes de los experimentos de campo. Las únicas dos limitantes incluyen el que sea posible o no la manipulación de una o más variables independientes, y al que las exigencias prácticas de la situación de investigación sean tales que el experimento de campo pueda realizarse sobre el problema particular bajo estudio. Superar estas dos dificultades no resulta fácil. Cuando puede hacerse, un amplio rango de problemas prácticos y teóricos se abren a la experimentación.

Como se indicó antes, las principales debilidades de los experimentos de campo son de índole práctica. La manipulación de las variables independientes y la aleatorización son quizás los dos problemas más importantes; son particularmente agudos en la investigación de ambientes escolares. La manipulación, aunque muy posible, a menudo puede no ser practicable porque, por ejemplo, los padres objetan cuando sus hijos, quienes fueron asignados aleatoriamente a un grupo control, no obtendrán un tratamiento experimental deseado. O quizá haya objeciones a un tratamiento experimental porque prive a los niños de alguna gratificación o los ubique en situaciones conflictivas.

No existe una razón real de por qué la aleatorización no pueda ser utilizada en los experimentos de campo. Sin embargo, con frecuencia se enfrentan dificultades. La mala voluntad para separar grupos de clases o para permitir que los niños sean asignados aleatoriamente a grupos experimentales son algunos ejemplos. Aun cuando la asignación aleatoria sea posible y permitida, la variable independiente puede empañarse seriamente a causa de que los efectos de los tratamientos no puedan aislarse de otros efectos. Por ejemplo, los maestros y los niños pueden discutir sobre lo que está sucediendo durante el curso del experimento. Para prevenir dicho oscurecimiento de las variables, el experimentador debería explicar a los administradores y maestros la necesidad de la asignación aleatoria y del control cuidadoso.

Una característica del campo experimental de naturaleza diferente es una debilidad en algunos experimentos, y en otros representa una fortaleza. Los investigadores de campo deben ser, por lo menos en cierto grado, operadores hábiles en lo referente a las habilidades sociales. Deben ser capaces de trabajar, hablar y convencer a la gente de la importancia y necesidad de su investigación. Deben estar preparados para pasar muchas horas, inclusive días y semanas, en paciente discusión con gente responsable de la situación institucional o comunitaria en la que van a trabajar. Por ejemplo, si van a trabajar en un sistema escolar rural, deben tener conocimientos sobre los problemas educativos rura-

les y generales, y sobre el sistema rural particular que desean estudiar. Algunos investigadores se tornan impacientes con estos aspectos preliminares, debido a que están ansiosos por realizar su trabajo de investigación. Encuentran difícil dedicar el tiempo y el esfuerzo necesarios en la mayoría de las situaciones prácticas. Otros disfrutan la socialización inevitable que acompaña a la investigación de campo. En French (1953) se presentan buenos consejos para el manejo de este aspecto de las situaciones de campo.

Un obstáculo importante para el buen diseño y que generalmente parece que se pasa por alto, es la actitud del investigador. Por ejemplo, la planeación de la investigación educativa parece caracterizarse con frecuencia por una actitud negativa resumida por afirmaciones tales como "eso no puede realizarse en escuelas", "los administradores y maestros no lo permitirán" y "no es posible hacer experimentos sobre este problema en esa situación". Iniciar con actitudes como éstas compromete cualquier buen diseño de investigación desde antes de que inicie la investigación. Si un diseño de investigación requiere de la asignación aleatoria de maestros a las clases, y si la falta de dicha asignación pone seriamente en riesgo la validez interna del estudio propuesto, debe realizarse cualquier esfuerzo para asignar aleatoriamente a los maestros. Los educadores que planean una investigación parecen suponer que los administradores o los maestros no permitirán el empleo de la asignación aleatoria. Sin embargo, esta suposición no es necesariamente correcta.

A menudo el consentimiento y cooperación de los maestros y administradores puede obtenerse si se utiliza un método apropiado, con orientación adecuada y precisa, y si se ofrecen explicaciones de las razones del uso de métodos experimentales específicos. Los puntos a enfatizar son los siguientes: diseñar investigaciones para obtener respuestas válidas a las preguntas de investigación; entonces, si es necesario hacer posible el experimento, se modifica el diseño "ideal". Con imaginación, paciencia y cortesía, muchos de los problemas prácticos de la implementación de un diseño de investigación pueden resolverse de manera satisfactoria.

Otra de las debilidades inherentes a las situaciones experimentales de campo es la falta de precisión. En el experimento de laboratorio es posible lograr un alto grado de precisión o exactitud, de tal manera que los problemas de medición y control de laboratorio generalmente son más simples que los de los experimentos de campo. En situaciones reales, siempre existe una gran cantidad de ruido sistemático y aleatorio. Para medir el efecto de una variable independiente sobre una variable dependiente en un experimento de campo, no es necesario únicamente maximizar la varianza de la variable manipulada ni de cualesquiera variables asignadas, sino también medir la variable dependiente de la manera más precisa posible. Pero en las situaciones reales, como en escuelas y grupos comunitarios, abundan las variables independientes extrañas. Además, las medidas de las variables dependientes, por desgracia, algunas veces no son lo suficientemente sensibles para recoger los mensajes de las variables independientes. En otras palabras, las medidas de la variable dependiente a menudo son tan inadecuadas que no pueden captar toda la varianza que las variables independientes han producido.

Estudios de campo

Los *estudios de campo* son investigaciones científicas no experimentales que buscan descubrir las relaciones e interacciones entre variables sociológicas, psicológicas y educativas en estructuras sociales reales. En este libro cualquier estudio científico (grande o pequeño), que busque relaciones de manera sistemática y que pruebe hipótesis, que sea no experimental y que se realice en situaciones de la vida (por ejemplo, comunidades, escuelas, fábricas, organizaciones e instituciones) será considerado un estudio de campo.

El investigador de un estudio de campo busca primero una situación social o institucional, y después estudia las relaciones entre las actitudes, valores, percepciones y conductas de individuos y grupos en dicha situación. El investigador del estudio de campo por lo común no manipula variables independientes. Antes de discutir y evaluar los diferentes tipos de estudios de campo sería útil considerar algunos ejemplos. Ya se han examinado estudios de campo en capítulos previos y en este capítulo se revisó el estudio Bennington de Newcomb. Ahora se examinarán brevemente dos estudios de campo más pequeños.

Anderson, Warner y Spencer (1984) estudiaron el sesgo por exageración de solicitantes de empleo. Los participantes de tal estudio eran solicitantes reales de puestos en el estado de Colorado. Los solicitantes de un empleo a menudo declaran tener más experiencia y más conocimientos de los que en realidad poseen. Para medir el grado de esta exageración, Anderson y sus colaboradores inventaron una actividad inexistente y les preguntaron a los solicitantes qué tanta experiencia tenían en dicha actividad. Los resultados demostraron que cerca de la mitad de los solicitantes declararon tener experiencia en una o más actividades inexistentes. Aquellos solicitantes que declararon tener gran experiencia en actividades inexistentes también exageraron su habilidad en tareas reales. Este estudio de campo ofrece información importante para aquellos involucrados en la toma de decisiones de contratación. Se trató de un estudio de campo pues no hubo una variable independiente manipulada. Se mencionaron actividades reales y falsas en un cuestionario y se les pidió a los participantes indicar la cantidad de experiencia que tenían en cada tarea, utilizando una escala de 4 puntos. Observe que dicho estudio no fue realizado en el laboratorio, y que utilizó participantes que no sospechaban la situación.

La investigación de campo realizada por Tom y Lucey (1997) estudió el tiempo de espera en las cajas de los supermercados y la satisfacción del cliente con el cajero y con la tienda. Estos investigadores estudiaron a cajeros rápidos y lentos durante periodos concurridos y no concurridos de las operaciones de la tienda. Los investigadores registraron los tiempos de espera de cada cliente y también entrevistaron al cliente cuando salía de la tienda. Los resultados demostraron que generalmente los clientes estaban más satisfechos respecto a la tienda y al cajero cuando el tiempo de espera percibido era más corto. Sin embargo, Tom y Lucey notaron que éste no era siempre el caso. En una de las dos tiendas utilizadas para el estudio, encontraron que algunos clientes reportaron mayor satisfacción con los cajeros lentos. Un cuestionamiento posterior reveló que los cajeros eran más lentos debido a que se tomaban el tiempo para dar al cliente una atención más personal.

Observe que los problemas de estos estudios de campo fueron atacados de manera no experimental: ni la aleatorización ni la manipulación experimental eran posibles. En el estudio de Jones y Cook, los datos fueron recolectados directamente de los estudiantes en dos universidades. En el estudio de Tom y Lucey sólo se utilizaron dos tiendas de abarrotes. Ninguno de estos estudios tuvo aleatorización o una variable independiente activa; no obstante, ambos fueron capaces de proporcionar información útil.

Tipos de estudios de campo

Katz (1953) dividió los estudios de campo en dos grandes tipos: *exploratorios* y de comprobación de hipótesis. El primer tipo, dice Katz, *busca lo que es*, en lugar de *predecir relaciones que se buscan*. El célebre estudio *Equality of Educational Opportunity*, citado en el capítulo 23, ejemplifica este tipo de estudio de campo. Los estudios exploratorios tienen tres propósitos: descubrir variables significativas en la situación de campo, descubrir relaciones entre variables y establecer las bases para una comprobación de hipótesis posterior, más sistemática y rigurosa.

Hasta este punto, en el libro se ha enfatizado el uso y la comprobación de hipótesis. Sin embargo, resulta conveniente reconocer que existen actividades preliminares a la comprobación de hipótesis en la investigación científica. Para lograr la meta deseada de la comprobación de hipótesis, con frecuencia debe realizarse investigación metodológica y de medición preliminar. Parte del mejor trabajo del siglo XX se ha realizado en esta área. Un ejemplo es el que lleva a cabo el analista factorial que está preocupado por el descubrimiento, aislamiento, especificación y medición de las dimensiones subyacentes del rendimiento, inteligencia, aptitudes, actitudes, situaciones y características de personalidad.

El segundo subtipo de estudios exploratorios de campo —investigación cuya meta es descubrir o revelar relaciones— es indispensable para el avance científico en las ciencias sociales. Es necesario conocer, por ejemplo, los correlatos de las variables. De hecho, el significado científico de un constructo surge de las relaciones que tienen con otros constructos. Suponga que no se tiene conocimiento científico del constructo "inteligencia"; no se conoce nada sobre sus causas o concomitantes. Por ejemplo, considere que no se sabe nada absolutamente sobre la relación entre inteligencia y rendimiento. Es concebible que se realice un estudio de campo en situaciones escolares. Podría observarse cuidadosamente un número de niños y niñas considerados inteligentes o no inteligentes por los maestros (aunque aquí se introduce contaminación ya que los maestros deben juzgar la inteligencia, por lo menos en parte, por el rendimiento). Quizá se observe que un gran número de los niños "más inteligentes" proviene de hogares de los niveles socioeconómicos más altos; que en clase resuelven problemas más rápido que otros niños; que poseen un vocabulario más amplio, etcétera. Ahora se tienen algunas pistas sobre la naturaleza de la inteligencia, de manera que es posible intentar construir una medida simple de inteligencia. Note que aquí la "definición" de inteligencia surge de lo que los supuestos niños inteligentes y no inteligentes hacen. Un procedimiento similar puede llevarse a cabo con la variable "rendimiento".

Fortalezas y debilidades de los estudios de campo

Los estudios de campo son fuertes en su realismo, significancia, fortaleza de las variables, orientación teórica y calidad heurística. La varianza de muchas variables en escenarios de campo reales es grande, especialmente comparada con la varianza de las variables de experimentos de laboratorio. Considere el contraste entre el impacto de normas sociales en un experimento de laboratorio como el de Sherif (1963), y el impacto de estas normas en una comunidad donde, por ejemplo, se aprueban ciertas acciones de los maestros y se desaprueban otras. Considere además la diferencia entre el estudio de la cohesión en el laboratorio, donde se le pregunta a los participantes, por ejemplo, si desean permanecer en un grupo (medida de cohesión), y el estudio de la cohesión del profesorado de una escuela, donde la permanencia en el grupo es parte esencial del futuro profesional de la persona. Compare la atmósfera de grupo en el estudio del Bennington College y la de un experimento de campo donde los instructores universitarios, que juegan diferentes papeles, crean diferentes atmósferas. Variables tales como clase social, prejuicio, conservadurismo, cohesión y clima social llegan a tener efectos fuertes en estos estudios. Sin embargo, la fortaleza de las variables no es una bendición pura. En una situación de campo por lo común existe tanto ruido en el canal que aunque los efectos sean fuertes y la varianza sea grande, para el experimentador no resulta fácil separar las variables.

El realismo de los estudios de campo es obvio. De todos los tipos de estudios, éstos son los que se asemejan más a la vida real; no puede haber una queja de artificialidad aquí. (Las observaciones sobre el realismo en los experimentos de campo se aplican, *a fortiori*, al realismo de los estudios de campo.)

Los estudios de campo son altamente heurísticos y *ad hoc*. Una de las dificultades de investigación de un estudio de campo es mantenerlo dentro de los límites del problema. Las hipótesis a menudo se le ocurren al investigador de inmediato debido a que el campo es rico en su potencial de descubrimiento. Por ejemplo, quizá se desee probar la hipótesis de que las actitudes sociales de los miembros del consejo de educación es un factor determinante de las decisiones políticas del consejo de educación. No obstante, después de empezar a reunir los datos, surgen muchos conceptos interesantes que pueden desviar el curso de la investigación.

A pesar de estas fortalezas, el estudio de campo es un primo científicamente débil de los experimentos de laboratorio y de campo. Su debilidad más seria, por supuesto, es su carácter no experimental. Por lo tanto, las proposiciones de relaciones son más débiles de lo que son en la investigación experimental. Para complicar las cosas, la situación de campo casi siempre tiene una gran cantidad de variables y de varianza. Piense en las muchas variables independientes posibles que pueden elegirse como determinantes de la delincuencia o del rendimiento escolar. En un estudio experimental dichas variables pueden ser controladas en gran parte, pero en un estudio de campo deben ser controladas, de alguna manera, con medios más indirectos y menos satisfactorios.

Otra debilidad metodológica es la carencia de precisión en la medición de variables de campo. Naturalmente, el problema de la precisión es más agudo en los estudios de campo que en los experimentos de campo. La dificultad encontrada por Astin (1968) para medir el ambiente universitario es uno de muchos ejemplos similares. Por ejemplo, se midió el ambiente administrativo a través de las percepciones de los estudiantes sobre los aspectos del ambiente. Mucha de la falta de precisión se debe a la mayor complejidad de las situaciones de campo.

Los estudios de organizaciones, por ejemplo, en su mayoría son estudios de campo, y la medición de las variables organizacionales ilustran bien las dificultades. "Efectividad organizacional" parece tan complejo como "efectividad del maestro". Para un análisis más profundo e ilustrativo véase Katz y Kahn (capítulo 8, 1978). Vale la pena una cuidadosa lectura y estudio de este excelente libro.

Otra debilidad de los estudios de campo incluye problemas prácticos: viabilidad, costo, muestreo y tiempo. Tales dificultades en realidad son debilidades *potenciales* —ninguna es necesariamente una debilidad real—. Las preguntas más obvias que se plantean son: ¿puede realizarse el estudio con las facilidades de que dispone el investigador? ¿Pueden medirse las variables? ¿Costará demasiado? ¿Requerirá de demasiado tiempo y esfuerzo? ¿Serán cooperadores los participantes? ¿Es posible el muestreo aleatorio? Cualquiera que contemple la posibilidad de un estudio de campo debe plantear y responder estas preguntas. Al diseñar la investigación es importante no subestimar las enormes cantidades de tiempo, energía y habilidad necesarios para la realización exitosa de la mayoría de los estudios de campo. El investigador de campo necesita ser un vendedor, administrador y empresario, a la vez que investigador.

Investigación cualitativa

Un área dentro de los estudios de campo es la investigación cualitativa. Hasta ahora, se ha hablado exclusivamente de investigación cuantitativa. Los estudios de campo con un énfasis cuantitativo poseen los problemas mencionados en la última sección. Sin embargo, la investigación cualitativa es diferente, pues no se basa en el uso de números o mediciones. Esta área de investigación cualitativa va creciendo en interés principalmente debido a que los investigadores se han dado cuenta de que no todos los estudios pueden o deben ser

cuantificados. Existen áreas de investigación donde los métodos cuantitativos no son capaces de captar adecuadamente la información. Por ejemplo, la investigación cuantitativa sería incapaz de captar información valiosa que sirviera para entender las experiencias de vida de pacientes renales que están bajo diálisis. La investigación cuantitativa puede proporcionar a los doctores y enfermeras información sobre la relación entre factores clínicos (tales como nutrición) y medidas de resultados (tales como tasas de supervivencia); pero no pueden indicar lo que el paciente en diálisis experimenta. Es la descripción de estas experiencias lo que permite el desarrollo de mejores programas de rehabilitación. El término "investigación cualitativa" se utiliza aquí para referirse a la investigación social y conductual basada en observaciones de campo discretas que se analizan sin utilizar números o estadística. Anteriormente se mencionó que quienes están involucrados con el aprendizaje operante o investigación skinneriana tampoco están interesados en el uso de la estadística inferencial; no obstante, se apartan de la investigación cualitativa, ya que sí utilizan números y mediciones. Los participantes de la investigación cualitativa pueden no estar conscientes de ser observados o estudiados. Varía el grado en que el participante se involucra en el proceso de investigación. A diferencia de la investigación de un solo sujeto o de series de tiempo, el participante no está consciente de que se estén haciendo mediciones. Dooley (1995) presenta un ejemplo sobresaliente de investigación cualitativa con el estudio de la teoría de la disonancia cognoscitiva. Dooley cita la investigación de Festinger (1956), quien estudió a la gente que predijo el fin del mundo pero que no vio su predicción convertirse en realidad. Este tipo de investigación requiere una metodología que no sea cuantitativa y que sea discreta. Sería muy difícil hacer que estas personas, que pertenecen a una secta, vinieran a un laboratorio de una universidad para ser estudiados. El investigador no tiene oportunidad de estudiar con eficacia a estas personas que acaban de experimentar disonancia cognoscitiva, pidiéndoles que completen un cuestionario o que participen en una entrevista estructurada. En su lugar, el investigador debe ser lo más discreto posible: tendría que actuar como alguien curioso o preocupado, e inclusive podría unirse a la secta como observador y encontrar la información requerida, de forma no amenazante. Se estudia a los participantes sin que ellos noten que están siendo estudiados. Sin embargo, Festinger realizó la investigación hace muchos años (1956). En el ambiente actual sería demasiado peligroso para los investigadores unirse a una secta con el propósito de estudiarla. ¿Por qué? En años recientes, especialmente en 1997, todos los miembros de una secta llamada Heaven's Gate se suicidaron por la llegada del cometa Hale-Bopp. Los miembros masculinos de esta secta estuvieron sujetos a severas alteraciones físico-quirúrgicas. Existe también una cantidad de sectas poderosas que utilizan métodos de programación rigurosos y drogas hipnóticas con sus miembros, para mantenerlos bajo control. Por lo tanto, aunque el ejemplo de Dooley es una buena ilustración de la investigación cualitativa, los autores de este libro de texto no recomiendan a alguien interesado en realizar investigación cualitativa sobre las sectas, unirse o convertirse en miembro de una de ellas.

Sería un poco más seguro considerar un estudio realizado por Rosenhan (1973), quien estaba interesado en la forma en que los hospitales psiquiátricos efectuaban diagnósticos especializados y en cómo serían las experiencias de un paciente psiquiátrico. Rosenhan pidió a ocho de sus cómplices que actuaran como pacientes psiquiátricos que sufrían de alucinaciones. Cada uno de estos seudopacientes fue admitido en diferentes hospitales. Durante su estancia, los seudopacientes nunca exhibieron algún síntoma. Los cómplices de Rosenhan realizaron observaciones sobre las condiciones hospitalarias, sobre cómo eran tratados y sobre el comportamiento del personal y de otros pacientes. Rosenhan reportó que el personal del hospital nunca supo que los seudopacientes no estaban enfermos.

No obstante, también existen estudios de investigación cualitativa donde el participante sabe que está participando en un estudio. En estos casos, el investigador necesita

desarrollar un alto nivel de empatía con los participantes. Por ejemplo, Jones (1998) utilizó el modelo cualitativo para estudiar a una cultura única (las bandas de adolescentes) en la sociedad estadounidense. Poco se ha reportado acerca de las bandas, con excepción de estadísticas. Se sabe poco sobre las dinámicas dentro de las bandas y sobre las diferencias entre algunos tipos de bandas. Jones tuvo que pasar una gran cantidad de tiempo en prisiones y centros de detención, entrevistando a miembros de bandas. Las experiencias de estar en una banda, las dinámicas entre sus miembros, sus sistemas de valores y cómo estos miembros de la sociedad estadounidense dan significado a sus vidas, satisfacen la meta de la metodología de la investigación cualitativa, la cual, como el estudio de Jones, resulta adecuada para estudiar experiencias de vida complejas.

La investigación cualitativa constituye un estudio de campo porque se realiza en el campo donde los participantes se comportan de manera natural. Heppner, Kivlighan y Wampold (1992) se refieren a la investigación cualitativa como naturalista-etnográfica o fenomenológica. Heppner y sus colaboradores presentan cuatro diferencias entre la investigación cualitativa y la cuantitativa (resumidas en la tabla 24.1).

La investigación cualitativa posee varias ventajas sobre la investigación cuantitativa. La primera utiliza observación directa y entrevistas semiestructuradas en escenarios del mundo real. El investigador busca transacciones e interacciones sociales entre la gente y los eventos. El proceso de recolección de datos resulta menos estructurado que en la investigación cuantitativa. El investigador puede hacer una serie de ajustes durante las observaciones; inclusive puede desarrollar nuevas hipótesis durante el proceso de investigación. La investigación cualitativa es naturalista, participativa e interpretativa.

La investigación cuantitativa rara vez se desvía del plan de investigación. La investigación cualitativa, por otro lado, es muy flexible. Esto ha provocado cierta crítica contra la investigación cualitativa. Algunos consideran que la investigación cualitativa sufre de algunos de los mismos problemas de validez, inherentes a los diseños de un solo sujeto. Otra área vulnerable es el sesgo del experimentador. El investigador cualitativo debe ser sumamente cuidadoso para evitar percibir las situaciones con un sesgo personal. No obstante, los investigadores cualitativos sostienen que el involucramiento discreto y la mezcla natural del observador con el ambiente reduce la cantidad de interrupción en el escenario y en el grupo de estudio. Después de un corto periodo, los participantes regresan a su forma normal de comportamiento y ya no muestran una *fachada*. El observador bien entrenado puede obtener percepciones del comportamiento de los participantes desde distintos pun-

TABLA 24.1 *Cuatro diferencias entre la investigación cuantitativa y la investigación cualitativa (Heppner, Kivlighan y Wampold)*

Cuantitativa	Cualitativa
Emana de la tradición post-positivista; los principales constituyentes son los objetos físicos y los procesos	Emana de la perspectiva fenomenológica; enfatiza eventos mentales internos como la unidad básica de la existencia
Asume que el conocimiento proviene de observaciones del mundo físico	El conocimiento se construye activamente y proviene del examen de los constructos internos de las personas
El investigador infiere, con base en observaciones directas o derivadas de observaciones directas	El investigador se basa en esquemas observacionales externos e intenta mantener intacta la perspectiva de los participantes
La meta consiste en describir causa y efecto	Intenta describir las formas en que la gente da significado al comportamiento

tos de vista. Si se realiza de manera apropiada, los datos recolectados por medio de la investigación cualitativa llegan a generar más información y menos variabilidad espuria que otros métodos de investigación. Quizá las dos visiones de la ciencia presentadas por Sampson (1991) en el capítulo 1 de este libro, incluyan las diferencias entre la investigación cuantitativa y cualitativa. En la investigación cualitativa la determinación del tamaño de la muestra puede realizarse cerca del final del estudio, en lugar de hacerlo al inicio, lo cual no es tan importante para el investigador cualitativo. Una regla de la investigación cualitativa es que a mayor número de entrevistas con cada participante, habrá menor necesidad de tener más participantes.

El diseño de la investigación cualitativa generalmente utiliza un observador discreto o un observador participante. Como observador discreto, el investigador realiza observaciones pasivas e intenta evitar responder al participante de cualquier manera. No se manipulan variables; el investigador sólo deja que los eventos naturales ocurran. Si el investigador deseara ver si la presencia de otra persona en el cuarto de baño afecta la disposición de alguien para lavarse las manos, entonces el investigador debe esperar y observar el comportamiento de la gente cuando haya otra persona en el cuarto de baño, y cuando no haya otra persona. En la investigación cuantitativa, el investigador utilizaría un cómplice para alterar la situación (véase Pedersen, Keithly y Brady, 1986). En la situación participante-observador, el investigador se vuelve parte del ambiente en estudio. Una característica de la forma participante-observador es que el investigador puede ver el efecto de manipular su propio comportamiento; de esta manera, en ocasiones los estudios de investigación cualitativa asemejan experimentos naturales.

Uno de los estudios de investigación cualitativa más famosos es el trabajo de Margaret Mead, quien estudió la cultura de Samoa. Dichos estudios no sólo se basan en observaciones personales sino que también requieren frecuentemente del reclutamiento de informantes. Estudios tales como los realizados para saber cómo era la vida de inmigrantes de primera generación, que llegaron a Estados Unidos en la primera parte del siglo XX, pueden ser estudios cualitativos. Los investigadores entrevistan a una cantidad de inmigrantes de primera generación y desarrollan historias de vida. Con suficientes historias de vida que muestren patrones de comportamiento similares, es posible desarrollar una descripción de cómo era la vida de quienes vivieron en esa época. Para mejores resultados, las entrevistas son videograbadas. El proceso de entrevista se conduce de tal manera y con tal duración que permite al informante adaptarse al entrevistador y al aparato de grabación. Es parte del plan de la investigación cualitativa elegir cuidadosamente al entrevistador para lograr la mejor combinación con el informante.

Puesto que los diarios, las grabaciones y las descripciones se obtienen de la gente estudiada en su ambiente natural, las cuestiones éticas son muy importantes; en particular, la confidencialidad de los registros y de la información debe mantenerse estrictamente segura. Hertz e Imber (1993) manifestaron que la investigación en ciencias sociales tiende a concentrarse en aquellos con menor poder (por ejemplo, animales, estudiantes universitarios) puesto que son de fácil acceso; mientras que los individuos poderosos no lo son (por ejemplo, políticos, ejecutivos corporativos, administradores escolares). Es poco probable que se le permita a un estudiante hacer un estudio de caso con el rector de una universidad. Por lo tanto, algunos estudios de investigación cualitativa incluyen el uso del engaño, lo cual es un tema que requiere de una revisión y justificación de cada caso.

Una excelente referencia sobre investigación cualitativa es la de Taylor y Bogdan (1998). Este libro está ya en su tercera edición y proporciona detalles claros del diseño, recolección de datos y el reporte final de la investigación cualitativa. Otra referencia muy útil sobre la investigación cualitativa es Cresswell (1998), quien señala que existen cinco tradiciones diferentes dentro de la investigación cualitativa. Él compara y critica la biografía, la

fenomenología, la teoría básica, la etnografía y el estudio de caso. Taylor y Bogdan, y Cresswell proporcionan ejemplos detallados de investigación cualitativa. Una excelente referencia sobre el uso de la metodología de investigación cualitativa para el estudio de pacientes con insuficiencia renal es *The Renal Rehabilitation Report*,[1] publicado por el Life Options Rehabilitation Advisory Council. En el volumen de julio/agosto de 1998 de esta publicación, se presenta una comparación entre el modelo tradicional y el modelo cualitativo. Tal artículo explica las razones por las que los métodos cualitativos son científicamente recomendables. Mientras Cresswell compara cinco diferentes tradiciones dentro de la investigación cualitativa, el artículo proporciona las descripciones de siete áreas diferentes. Entre las categorías se encuentran la investigación feminista, la investigación de acción y la investigación de evaluación cualitativa. La investigación feminista se enfoca en la mejora de las necesidades, intereses, experiencias y metas de las mujeres. La investigación de acción implica el esfuerzo conjunto del investigador y del participante para lograr un cambio. La investigación de evaluación cualitativa trata con historias y estudios de caso.

A pesar de que se presentó una visión positiva de los métodos de investigación cualitativa, no todos los individuos mantienen la misma opinión. La mayoría de las ciencias del comportamiento —especialmente la psicología— se han mostrado a favor del modelo cuantitativo. Existen algunos, como Sampson (1991) y Phillips (1973), quienes han afirmado que la cuantificación no es un método apropiado para todas las situaciones de investigación. Ya se analizaron brevemente los beneficios de los métodos cuantitativos; sin embargo, al mismo tiempo, se mencionó que son incapaces de responder ciertas preguntas respecto a la cultura o a ciertas formas de vida. Este conflicto entre la metodología de la investigación cualitativa y la metodología cuantitativa se encuentra bien documentado en la literatura (véase Cook y Reichardt, 1979; Padgett, 1998). Existen fieles defensores de ambas posturas; sin embargo, los investigadores cuantitativos, como Cook y Reichardt, han discutido los temas y presentado algunas ideas para combinar ambas, en lugar de separarlas. Ellos hablan acerca de la posibilidad de un estudio de investigación que combine elementos cuantitativos y cualitativos. De hecho, Padgett (1998) describe tres formas para realizar tanto investigación cuantitativa como cualitativa en un estudio. La combinación de los dos métodos —cuantitativo y cualitativo— se llama *investigación multimétodo*.

Según Padgett, la primera de las tres formas de hacer investigación multimétodo consiste en iniciar la investigación de manera cualitativa y terminarla de manera cuantitativa. El método cualitativo sirve para explorar e identificar las ideas, hipótesis y variables de interés para el investigador. Esto se haría por medio de observación directa, entrevistas o grupos de enfoque. Los conceptos derivados de la porción cualitativa del estudio pueden, entonces, estudiarse a través del uso de métodos cuantitativos y de la comprobación de hipótesis. La generalización de los conceptos y de las hipótesis, probados a través de la investigación cuantitativa, puede proporcionar mayor credibilidad al obtener un mejor vínculo con el mundo real. Los métodos cualitativos proporcionarían ese vínculo.

La segunda forma de hacer investigación multimétodo es utilizar el método cuantitativo primero, seguido por el método cualitativo. Los resultados de la porción cuantitativa del estudio se usan como el punto de inicio de la porción cualitativa. Padgett considera que muchos estudios cuantitativos podrían beneficiarse de un análisis cuantitativo de los resultados. Los métodos cualitativos pueden ayudar a proporcionar entendimiento e información respecto a las preguntas que no fueron respondidas o no podían responderse por medio del estudio cuantitativo. Por ejemplo, en estudios cuantitativos que utilizan el

[1] Una copia de este reporte está disponible en el Life Options Rehabilitation Resource Center al (800) 468-7777. Los autores desean agradecer al doctor Abdul Abukurah por proporcionarnos una copia de esta publicación.

análisis de regresión múltiple (que se cubre en un capítulo posterior de este libro), el investigador a menudo se queda con un cierto porcentaje de varianza injustificada. Por ejemplo, un estudio de investigación que reporta un coeficiente de correlación de .48, entre las puntuaciones del Graduate Record Examination (GRE) y el éxito en los estudios de posgrado, está indicando que únicamente el 23% de la varianza total del éxito en los estudios de posgrado se explica por las puntuaciones del GRE. Esto también indica que el 77% no está justificado. En este punto, a través del uso de los métodos cualitativos, se puede iniciar el proceso de determinar qué otras variables estarían involucradas. Esto puede, a su vez, conducir a otro estudio cuantitativo que incluya aquellas variables encontradas en la porción cualitativa del estudio.

La tercera forma descrita por Padgett difiere ligeramente de las dos primeras, ya que tiene una división temporal más definida; es decir, después de completar un método, continúa el otro. En la tercera forma de investigación multimétodo, ambos métodos, cualitativo y cuantitativo, se utilizan simultáneamente. Dichos métodos pueden tener un método más dominante que el otro. Cuando ello sucede, un método —el menos dominante— se "anida" dentro del otro —el dominante—. Padgett informa que existen más estudios con esta naturaleza "anidada" que la integración verdadera de los dos métodos en el estudio. En el caso en que los investigadores continúan un resultado cuantitativo con un hallazgo cualitativo, se dice que los métodos cualitativos complementan pero no alteran el modelo cuantitativo del estudio. En el caso opuesto, donde el método cualitativo es el dominante, el investigador realiza una encuesta o entrevista; pero utiliza escalas e instrumentos de medición estandarizados en el proceso, lo cual incluye el uso de escalas de tipo Likert y datos de censo para complementar los datos obtenidos de las entrevistas intensivas. Aquí, los datos cuantitativos no se entrometen dentro de la naturaleza inductiva y holística de los métodos cualitativos.

A pesar de que el uso conjunto de los métodos cualitativos y cuantitativos resulta promisorio, existen todavía algunas dudas en las mentes de muchos. Cook y Reichardt (1979), por ejemplo, señalan algunos de los obstáculos que enfrenta la investigación multimétodo. Los obstáculos que señalan se relacionan principalmente con la economía y el entrenamiento. Un estudio con los esfuerzos conjuntos de los métodos cualitativo y cuantitativo puede ser costoso en términos de tiempo y dinero. Aun cuando la investigación multimétodo requiera de fe y vigilancia, Padgett (1998) considera que la investigación multimétodo vale el costo y el esfuerzo.

Anexo

El paradigma experimental holístico

El paradigma experimental holístico proporciona un medio económico de cuantificación empírica de las relaciones complejas entre los factores críticos que afectan el desempeño humano, sobre tareas operacionales individuales. El modelo produce una ecuación de orden requerido por la mayoría de los factores potencialmente críticos relacionados con la tarea, la gente, el equipo y el tiempo, a través de sus rangos operacionales efectivos. Al combinarse con varias técnicas de reducción del sesgo, el modelo holístico mejora materialmente la precisión predictiva de los resultados experimentales y produce información más generalizable de lo que permite el modelo de pocos factores a la vez.

Contrario a los alegatos efectuados en muchos de los libros de texto de las ciencias del comportamiento actuales, respecto a los experimentos factoriales grandes, el modelo es extremadamente económico. De hecho, con el uso de diseños fraccionales de manera

secuencial, resulta mucho menos costoso llevar a cabo experimentos megafactoriales,[2] de lo que sería obtener la información respecto al mismo número de factores en una serie de experimentos pequeños. Los experimentos megafactoriales grandes tampoco son meras extensiones de experimentos más pequeños; se realizan de manera diferente. Requieren de menos supuestos de los tipos presentes en los experimentos de pocos factores, y en el paradigma holístico los pocos supuestos que se hacen son tentativos y se probarán eventualmente conforme el experimento progrese, y se modifican conforme se necesite. Mientras que la metodología se adapta a los problemas que involucren factores cuantitativos y el modelo del ANOVA, muchos de sus principios pueden utilizarse en el camino, en investigación científica del comportamiento. El modelo es heurístico, pragmático y empírico.

El paradigma experimental holístico conforma una metodología completa, que integra un conjunto de principios, una estrategia y un cuerpo de técnicas que proporcionan respuestas cuantitativas con un sesgo mínimo a preguntas complejas del comportamiento. El sesgo se define como la diferencia entre los estimados del desempeño, basada en los resultados y realizaciones experimentales, obtenidas bajo condiciones operacionales. La estrategia básica en este modelo holístico fue tomada de la metodología G.E.P. de superficie de respuesta de Box (Box, 1954), modificada para ajustarla a los problemas especiales que se encuentran en los experimentos del comportamiento. Sin embargo, el modelo holístico no depende de un diseño experimental o técnica estadística específica; por el contrario, las disposiciones estadísticas juegan un papel mucho menor que en los experimentos tradicionales.

El modelo holístico enfatiza una planeación pre-experimental y una fase de exploración, como un punto para verificar las condiciones que afecten de forma negativa la conducción del experimento y el desempeño de los operadores. Al mismo tiempo, los factores experimentales relevantes se seleccionan con base en el conocimiento actual del investigador y con pruebas preliminares. Después, con el uso de la estrategia de Box, utilizando una secuencia de diseños factoriales fraccionales, esos factores se estudian al nivel más bajo de resolución, una ecuación de primer orden. A continuación, si se encuentra que este modelo no se ajusta adecuadamente a los datos empíricos, se recolecta otro grupo de datos para expandir el orden de la ecuación y se realiza otra prueba. Puesto que la mayor parte del desempeño humano puede ser aproximado adecuadamente por no más que un polinomio de tercer orden, por lo común se concluye este modelo iterativo, después de tomar mucho menos datos de los que se requerirían para llenar el diseño factorial.

Las técnicas más importantes empleadas en este modelo fueron desarrolladas principalmente en los años treinta y sesenta. Se desarrollaron nuevas técnicas para robustecer la recolección secuencial de datos de cientos de condiciones experimentales, para las tendencias y la transferencia intraserial o efectos "remanentes", sin compensación o aleatorización excesivas. Otras técnicas empleadas en el modelo holístico incluyen transformaciones gráficas y análisis gráfico de datos.

Un análisis crítico del modelo tradicional de la experimentación del comportamiento revela que muchos de sus ritos se han tornado sacrosantos, deificados en algo que no son. Conforme este libro pasa a impresión, se desafía uno de los iconos de la ciencia del comportamiento: la prueba de la significancia estadística (véase Harlow, Mulaik y Steiger, 1997), como ya ha sucedido a menudo durante más de 30 años (véase capítulo 1 de Bakan, 1973).

[2] El doctor Charles Simon acuñó este término para evitar una confusión con la palabra "multifactorial", a la que los escritores de libros de texto con frecuencia se refieren como experimentos de 2, 3 y 4 factores. La primera definición de "mega" es "grande" y "megafactor" implica un número mucho mayor de factores de aquellos que tradicionalmente se han utilizado.

Otras llamadas reglas de investigación científica han dictado experimentos que producen resultados de poco o ningún valor duradero y algunas ocasiones con conclusiones totalmente incorrectas. Esto sucede, por ejemplo, cuando factores críticos no incluidos en el experimento se mantienen constantes. La elección de los valores constantes en los que se mantienen dichos factores, llegan a alterar el nivel de dificultad de la tarea y alterar marcadamente los resultados generales. La aleatorización no garantiza evitar el sesgo ni garantiza la "validez interna", y únicamente debe utilizarse después de haber agotado todos los controles sistemáticos conocidos.

Un procedimiento, recomendado frecuentemente por los tradicionalistas, para incrementar la "validez externa" o generalización de los resultados del experimento consiste en realizar unos cuantos estudios con parámetros modestamente diferentes, después de completar el experimento principal. Éste es un costoso modelo de ensayo y error.

La generalización se logra con mayor precisión y menor costo en el modelo holístico al incluir todos los factores relevantes identificables en el plan experimental original.

El paradigma experimental holístico fue desarrollado por Charles W. Simon durante los pasados 30 años, apoyado principalmente por los departamentos de investigación de la fuerza aérea, la marina y el ejército de Estados Unidos. En los años setenta se impartieron seminarios a grupos de investigación industriales y militares. Hasta la fecha, no está disponible un reporte consolidado, sino únicamente numerosos reportes de las diferentes técnicas, a menudo aislados unos de otros, y no necesariamente actualizados con desarrollos recientes. Un libro está actualmente en preparación.[3]

RESUMEN DEL CAPÍTULO

1. Se comparan y contrastan los experimentos de laboratorio, los experimentos de campo y los estudios de campo.
2. El experimento de laboratorio posee la mayor validez interna, pero tiene debilidad en su validez externa.
3. Los experimentos de laboratorio por lo común muestran variables con un efecto pequeño; mientras que los estudios de campo y los experimentos de campo muestran variables con efectos grandes.
4. Aunque los experimentos de campo tienen variables que muestran efectos grandes, con frecuencia dicho efecto es enmascarado por otras variables siendo difícil superar este obstáculo.
5. Los experimentos de laboratorio tienen una gran orientación teórica, y se diseñan para probar una teoría general.
6. Los experimentos de campo y los estudios de campo tienen una mayor orientación aplicada, y buscan responder una pregunta específica sobre fenómenos observables.
7. Los experimentos de campo difieren de los experimentos de laboratorio, ya que los primeros no poseen los controles estrictos hallados en la investigación de laboratorio.
8. Los experimentos de campo intentan conducir un estudio del tipo del laboratorio en un ambiente del mundo real, con el uso de participantes del mundo real. Casi siempre hay una variable independiente activa.
9. Los estudios de campo son estudios no experimentales realizados en el mundo real. Por lo común no hay una variable independiente activa.

[3] El Dr. Charles W. Simon preparó esta descripción del modelo holístico.

10. La meta de los estudios de campo es descubrir las relaciones e interacciones entre un número de variables sociales y del comportamiento.
11. La mayoría de la investigación de las ciencias del comportamiento y de las ciencias sociales tiene una orientación cuantitativa. A un estudio de campo orientado cuantitativamente a menudo se le llama encuesta o investigación epidemiológica. A un estudio de campo orientado cualitativamente se le denomina investigación cualitativa o naturalista-etnográfica.
12. La investigación cualitativa incluye métodos con observadores discretos o con observadores participantes.
13. En el método del observador discreto, éste se mezcla con el ambiente y no tiene contacto con los participantes. El método del observador participante requiere que el observador se convierta en miembro del grupo en estudio.
14. Los métodos cualitativos son adecuados para estudiar experiencias humanas poco conocidas o complejas. La investigación cualitativa complementa la investigación cuantitativa y no pretende suplantarla.

SUGERENCIAS DE ESTUDIO

1. ¿Dónde es más probable el uso del análisis factorial de varianza, en los experimentos de laboratorio, en los experimentos de campo o en los estudios de campo? Explique.
2. En el capítulo 15 se describió un estudio sobre los efectos comparativos de la marihuana y el alcohol. Suponga que dicho estudio es un experimento de laboratorio. ¿Eso limita su utilidad y generalización? ¿Diferiría un estudio como éste, respecto a la generalización, de un experimento de laboratorio sobre frustración y agresión?
3. A continuación se presenta una lista de estudios. Algunos se resumen en capítulos previos y otros no. Consulte estos estudios y después clasifique cada uno como experimento de laboratorio, experimento de campo o estudio de campo. Explique por qué categoriza cada estudio de esa manera.

   Henemann, H. G. (1977). Impact of test information and applicant sex on applicant evaluation in a selection simulation. *Journal of Applied Psychology, 62*, 524-526.

   Johnson, C. B., Stockdale, M. S. y Saal, F. E. (1991). Persistence of men's misperceptions of friendly cues across a variety of interpersonal encounters. *Psychology of Women Quarterly, 15*, 463-475.

   McKay, J. R., Alterman, A. I., McLellan, T., Snider, E. C. y O'Brien, C. P. (1995). Effect of random versus nonrandom assignment in a comparison of inpatient and day hospital rehabilitation for male alcoholics. *Journal of Consulting and Clinical Psychology, 63*, 70-78.

   Reinholtz, R. K. y Muehlenhard, C. L. (1995). Genital perceptions and sexual activity in a college population. *Journal of Sex Research, 32*, 155-165.

   Wansink, B., Kent, R. J. y Hoch, S. J. (1998). An anchoring and adjustment model of purchase quantity decisions. *Journal of Marketing Research, 35*, 71-81.

   Wilson, F. L. (1996). Patient education materials nurses use in community health. *Western Journal of Nursing Research, 18*, 195-205.

4. "El experimento es uno de los más grandes inventos del último siglo." ¿Concuerda usted con esta afirmación? Si es así, mencione las razones para ello: ¿por qué es correcta la afirmación (si, de hecho, es correcta)? Si no está de acuerdo, explique por

qué no. Antes de formular juicios rápidos, lea y pondere las referencias que se ofrecen en las sugerencias de estudio número 6, abajo.

5. Por desgracia ha habido mucha crítica desinformada sobre los experimentos. Antes de realizar juicios racionales sobre cualquier fenómeno complejo se debe conocer primero sobre lo que se está hablando y, segundo, se debe conocer la naturaleza y el propósito del fenómeno que se critica. Para ayudarle a obtener conclusiones racionales acerca del experimento y de la experimentación, se ofrecen las siguientes referencias como lectura previa.

Berkowitz, L. y Donnerstein, E. (1982). External validity is more than skin deep: Some answers to criticisms of experiments. *American Psychologist, 3*, 245-257. (Una respuesta profunda a la crítica sobre la carencia de validez externa de los experimentos.)

Kaplan, A. (1964). *The conduct of inquiry*. San Francisco, California: Chandler. (El capítulo IV, llamado "Experiment", parece incluir la observación más controlada.)

6. Aquellos que deseen conocer más sobre el paradigma experimental holístico pueden consultar algunas de las primeras publicaciones de Charles W. Simon, que brindan la filosofía subyacente al modelo. Tome en cuenta que desde la época en que fueron escritos, se ha refinado el modelo y se le han incorporado nuevas técnicas.

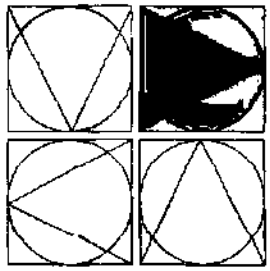
Simon, C. W. (1976). Analysis of human factors engineering experiments: Characteristics, results and applications. Westlake Village, California: Canyon Research Group, Inc., Tech. Rep. No. CWS-02-767, 104 pp. (AD A038-184).

Simon, C. W. (1978). New research paradigm for applied experimental psychology: A system approach. Westlake Village, California: Canyon Research Group, Inc., Tech. Rep. No. CWS-04-77A, 123 pp. (AD A 056-984).

Simon, C. W. (1987). Will egg-sucking ever become a science? *Human Factors Society Bulletin, 30*, 1-4.

Simon, C. W. y Roscoe, S. N. (1984). Application of a multifactor approach to transfer of learning research. *Human Factors, 26*, 591-612.

Westra, D. P., Simon, C. W., Collyer, S. C. y Chambers, W. S. (1982). *(Simulator design features for carrier landings: I. Performance experiments.* NAVTRAEQUIPCEN. (78-C-0060-7): 64 pp. (AD A122-064).



CAPÍTULO 25

INVESTIGACIÓN POR ENCUESTA

La investigación por encuesta estudia poblaciones (o universos) grandes o pequeñas, por medio de la selección y estudio de muestras tomadas de la población, para descubrir la incidencia, distribución e interrelaciones relativas de variables sociológicas y psicológicas. Como tal, la investigación por encuesta puede clasificarse como estudios de campo con una orientación cuantitativa. Algunos la consideran una variación del diseño de investigación correlacional. Este capítulo se concentra en el empleo de la investigación por encuesta en la investigación científica y rechaza las llamadas encuestas de estatus, las cuales tienen una meta diferente de la investigación por encuesta: su meta es conocer el *status quo*, en lugar de estudiar las relaciones entre variables; así como examinar la situación actual de algunas características poblacionales. Las investigaciones por encuesta se utilizaban ya en 1830 en Gran Bretaña, para estudiar las condiciones laborales de niños y adultos durante la Revolución Industrial. La investigación por encuesta en las ciencias sociales y del comportamiento es más moderna, pues es un desarrollo del siglo XX.

No existe la intención de ir en contra de las encuestas de estatus, ya que son útiles e, incluso, indispensables. La intención es destacar la importancia y utilidad de la investigación por encuesta en el estudio científico de problemas social y educativamente significativos. El trabajo de los encuestadores sobre opinión pública, tal como Gallup y Roper, no será examinado. Si se desea revisar una buena explicación sobre encuestas y sus tipos, véase Parten (1950), en el capítulo 1. (Aunque antiguo, este libro aún es valioso.) Un libro un poco más nuevo, sobre la forma en que se utilizan las encuestas para captar la opinión

pública en Estados Unidos, es el de Wheeler (1976). El texto clásico, utilizado durante muchos años, hasta que recientemente dejó de editarse, es el de Warwick y Lininger (1975); tiene la ventaja de haber sido guiado por las ideas y práctica del Survey Research Center de la University of Michigan. También posee la ventaja de tener un énfasis transcultural. Orlich (1978) presenta el método y el procedimiento de la investigación por encuesta de una forma muy directa. Inclusive explica cómo diseñar y secuenciar los reactivos de una encuesta. Algunas de las publicaciones más recientes sobre investigación por encuesta son las realizadas por Alreck (1994), Babbie (1990), Suskie (1996) y Weisberg (1996).

Las encuestas incluidas en la definición anterior con frecuencia se denominan *encuestas muestrales*, quizá debido a que la investigación por encuesta se desarrolló como una actividad de investigación separada, junto con el desarrollo y mejoramiento de los procedimientos de muestreo. La investigación por encuesta se considera como una rama de la investigación científica social, la cual se distingue inmediatamente de la encuesta de estatus. Sus procedimientos y métodos han sido desarrollados principalmente por psicólogos, sociólogos, economistas, científicos políticos y estadísticos (véase Campbell y Katona, 1953). Estos individuos han puesto un sello de rigor científico en la investigación por encuesta y, en el proceso, han influido profundamente en las ciencias sociales.

La definición también liga poblaciones y muestras. Los investigadores de encuestas están interesados en la evaluación precisa de las características de poblaciones completas de personas. Ellos desean saber, por ejemplo, cuántas personas en Estados Unidos votaron por un candidato republicano, y la relación entre dicha votación y variables como sexo, raza, preferencia religiosa y otras similares. Buscan conocer la relación entre las actitudes hacia la educación y el apoyo público a los presupuestos escolares.

Sin embargo, sólo en raras ocasiones los investigadores por encuestas estudian poblaciones completas; más bien estudian *muestras* obtenidas de poblaciones. A partir de las muestras ellos infieren las características de la población o universo definidos. El estudio de muestras, a partir de las cuales se pueden realizar inferencias sobre poblaciones, es necesario debido a las dificultades para estudiar poblaciones completas. Las muestras aleatorias pueden, frecuentemente, generar la misma información que un censo (una enumeración y estudio de una población entera), a un costo mucho menor, con mayor eficiencia y, algunas veces, ¡con mayor precisión!

Algunas encuestas intentan determinar la incidencia, distribución e interrelaciones entre variables sociológicas y psicológicas y, al hacerlo, generalmente se enfocan en la gente, los factores vitales de la gente, y sus creencias, opiniones, actitudes, motivaciones y comportamiento. La naturaleza científica social de la investigación por encuesta se revela por la naturaleza de sus variables, que pueden clasificarse como hechos, opiniones y actitudes sociológicas. Los *hechos sociológicos* son atributos de los individuos que surgen de su pertenencia a grupos sociales: sexo, ingreso, afiliaciones políticas y religiosas, nivel socioeconómico, educación, edad, gasto para vivir, ocupación, raza, etcétera. El segundo tipo de variable es psicológica e incluye *opiniones* y *actitudes*, por un lado, y *comportamiento* por el otro. Los investigadores por encuestas no se interesan únicamente en las relaciones entre variables sociológicas; tienden a interesarse más en lo que la gente piensa y hace, así como en las relaciones entre variables sociológicas y psicológicas. El estudio sobre la calidad de vida en Estados Unidos, realizado por el Survey Research Center de la University of Michigan, por ejemplo, ofrece datos deprimentes sobre la relación entre la raza y los sentimientos de confianza en la gente, una variable sociológica y psicológica (los datos se presentan en la tabla 25.1). La relación es sustancial. En efecto, las personas afroamericanas confían menos en la gente, que las personas blancas. Como Campbell, Converse y Rodgers (1976) afirman (p. 455), "aquellas personas que han tenido menos éxito en sus encuentros con la sociedad tienen menos razones para tener confianza en ella".

TABLA 25.1 *Relación entre raza y confianza en la gente (en porcentajes) (estudio de Campbell y colaboradores)*[a]

	Poca confianza	Mucha confianza
Afroamericanos	72	28
Americanos blancos	38	62

[a] $N = 2\,070$.

Por supuesto, los investigadores por encuesta también estudian las relaciones entre variables psicológicas; sin embargo, las relaciones de la investigación por encuesta ocurren entre variables sociológicas y psicológicas: entre educación y tolerancia, entre raza y autoestima y entre educación y sentido de eficacia política.

Tipos de encuestas

Las encuestas pueden ser clasificadas convenientemente de acuerdo con los siguientes métodos para obtener información: entrevista personal, cuestionario enviado por correo, por panel y por teléfono. De éstas, la entrevista personal eclipsa, por mucho, a las otras, y quizás sea la herramienta más poderosa y útil de la investigación social científica por encuesta. Estos tipos de encuestas se describirán brevemente; en capítulos posteriores, cuando se revisen los métodos de recolección de datos, se estudiará con profundidad la entrevista personal.

Entrevistas e inventarios

La mejor investigación por encuesta utiliza la entrevista personal como método principal para obtener información. Esto se logra, en parte, por la construcción cuidadosa y laboriosa de un inventario o cuestionario. Se utilizará el término "inventario" pues tiene un significado claro: es el instrumento utilizado para reunir información de encuesta, a través de una entrevista personal. El término "cuestionario" ha sido utilizado para nombrar instrumentos personales de entrevista e instrumentos de actitudes o de personalidad. Estos últimos se llaman "escalas" en este libro. La información del inventario incluye información factual, opiniones y actitudes, y razones del comportamiento, de opiniones y de actitudes. Los inventarios de entrevista son difíciles de construir, consumen mucho tiempo y son relativamente costosos; pero ningún otro método proporciona la información que ellos ofrecen.

La *información factual* obtenida en encuestas incluye los llamados datos sociológicos que se mencionaron antes: género, estado civil, educación, ingreso, preferencia política, preferencia religiosa y otros similares. Dicha información resulta indispensable, ya que sirve para estudiar las relaciones entre variables y para verificar la adecuación de las muestras. Estos datos, que se anotan en una "carátula", se denominan "información de la carátula". Esta información, o al menos parte de ella, casi siempre se obtiene al inicio de la entrevista. La mayoría de ella es de carácter neutral y ayuda al entrevistador a establecer empatía con el entrevistado. Las preguntas de naturaleza más personal, tales como los hábitos personales y el ingreso, y preguntas más difíciles de responder, tales como el nivel de conocimiento o habilidad del entrevistado, pueden reservarse para un cuestionamiento posterior, quizás para el final del inventario. Saber cuál es el momento adecuado debe ser necesariamente una cuestión de juicio y experiencia (véase Warwick y Lininger, 1975).

Otros tipos de información factual incluyen lo que los entrevistados saben sobre el tema de investigación, lo que ellos hicieron en el pasado, lo que están haciendo ahora y lo que pretenden hacer en el futuro. Después de todo, a menos que se observe de manera directa, todos los datos sobre el comportamiento de los entrevistados deben provenir de ellos o de otras personas. En este especial sentido, todo el comportamiento pasado, presente y futuro puede clasificarse bajo el "hecho" de comportamiento, aun cuando el comportamiento tan sólo sea una intención. Un aspecto importante de dichas cuestiones factuales es que el entrevistador presumiblemente conoce bastante sobre las acciones y comportamientos personales. Si el entrevistado dice que votó por una emisión de bonos escolares, puede considerarse que la afirmación es verdadera —a menos de que exista evidencia contundente de lo contrario—. De forma similar, puede creérsele al entrevistado, quizás con mayor reserva (puesto que el hecho aún no ha sucedido), si manifiesta su intención de votar por la emisión de los bonos escolares.

Las creencias, opiniones, actitudes y sentimientos que tienen los entrevistados acerca de los objetos cognitivos son de gran importancia, quizás aún más importantes desde un punto de vista científico social. *Objetos cognitivos* es una expresión que indica el objeto de una actitud. Casi cualquier cosa puede ser el objeto de una actitud, pero generalmente el término se reserva para "objetos" sociales importantes, por ejemplo, grupos (religiosos, raciales y educativos) e instituciones (escuela, matrimonio y partidos políticos). Un término más general y quizá más adecuado, aunque no de uso general, es *referente*. Muchos de los objetos cognitivos de la investigación por encuesta pueden no ser de interés para los investigadores: inversiones, ciertos productos comerciales, candidatos políticos y otros similares. Otros objetos cognitivos tal vez sean más interesantes: las Naciones Unidas, la Suprema Corte, las prácticas educativas, la integración, el comportamiento sexual, el subsidio federal a la educación, los estudiantes universitarios y el movimiento feminista.

La entrevista personal puede ayudar a conocer las razones del entrevistado para hacer o creer algo. Cuando se le pregunta a la gente las razones de las acciones, intenciones o creencias, llegan a responder que han hecho algo, intentado hacer algo o que se sienten de ciertas formas acerca de algo. Pueden decir que afiliaciones grupales, lealtades o ciertos eventos han influido en ellos; o pueden haber escuchado acerca de temas bajo investigación por los medios de comunicación masiva. Por ejemplo, un entrevistado masculino puede declarar que antes se oponía al subsidio federal para la educación a causa de que él y su partido político siempre se habían opuesto a la interferencia gubernamental. No obstante, ahora apoya la ayuda federal porque ha leído mucho acerca del problema en periódicos y revistas, y ha llegado a la conclusión de que el subsidio federal beneficiará a la educación del país.

Los deseos, valores y necesidades de un entrevistado influyen en sus actitudes y acciones. Al explicar por qué está de acuerdo con el subsidio federal a la educación, quizá manifieste que sus propias aspiraciones educativas se vieron truncadas y que siempre ha abogado por una mayor educación; o quizá señale que su grupo religioso tiene un fuerte compromiso con la educación infantil como parte de sus estructura de valores. Si el individuo bajo estudio ha revelado en forma concreta sus valores, deseos y necesidades —y llega a expresarlos de manera verbal— la entrevista personal resulta muy valiosa.

Otros tipos de investigación por encuesta

El siguiente tipo importante de la investigación por encuesta es el *panel*. Se selecciona una muestra de participantes y se les entrevista; después se entrevistan de nuevo y se estudian posteriormente. La técnica del panel permite que el investigador estudie cambios en

ambigua y la especificación del problema de investigación; así como el análisis e interpretación de los datos.

En el espacio limitado de una sección de un capítulo es, de hecho, imposible analizar adecuadamente la metodología de la investigación por encuesta. Por lo tanto, se incluirán sólo aquellas partes de la metodología relacionadas con los propósitos de este libro: el diseño de la encuesta o del estudio, el llamado plan o gráfica de flujo de los investigadores por encuestas, la verificación de la confiabilidad y la validez de la muestra, y los métodos de recolección de datos. (Tanto el muestreo como el análisis ya se estudiaron en capítulos previos.)

Los investigadores de encuestas utilizan un *plan* o *gráfica de flujo* para bosquejar el diseño y las implementaciones subsecuentes de una encuesta. El plan de flujo inicia con los objetivos de la encuesta, enlista cada paso a seguir y termina con el reporte final. Primero se establecen los problemas generales y específicos a resolver, de la forma más cuidadosa y completa posible. Puesto que, en principio, no hay nada muy diferente aquí del análisis de los problemas e hipótesis del capítulo 2, puede omitirse un estudio detallado y presentar un ejemplo hipotético simple. Un investigador educativo ha sido comisionado por un consejo de educación para estudiar las actitudes de los miembros de la comunidad hacia el sistema escolar. Al discutir el problema general con el consejo y con los administradores del sistema escolar, el investigador nota que hay varios problemas más específicos, tales como: ¿la actitud de los miembros de la comunidad se ve afectada por el hecho de tener a sus hijos en la escuela? ¿Su nivel educativo afecta sus actitudes?

Uno de los trabajos más importantes del investigador consiste en especificar y aclarar el problema. Para hacerlo bien, el investigador no debe preguntar a las personas únicamente lo que piensan de las escuelas, aunque ésta pueda ser una buena manera de empezar, en caso de no saber mucho sobre el tema. Deben plantearse preguntas específicas, orientadas a las diferentes facetas del problema. Cada una de estas preguntas debe incluirse en el inventario de la entrevista. Algunos investigadores de encuesta diseñan tablas para el análisis de los datos en este momento, para aclarar el problema de investigación y para guiar la construcción de las preguntas de la entrevista. Ya que dicho procedimiento es recomendable, se diseñará una tabla para demostrar cómo puede utilizarse para especificar los objetivos y las preguntas de la encuesta.

Considere la pregunta: ¿está relacionada la actitud con el nivel educativo? La pregunta requiere que la "actitud" y el "nivel educativo" sean definidos operacionalmente. Las actitudes negativas y positivas se inferirán a partir de las respuestas a las preguntas y reactivos del inventario. Si, en respuesta a una pregunta tan vaga como "en general, ¿qué piensa de nuestro sistema escolar?", el encuestado responde: "es uno de los mejores en esta área", puede inferirse que tiene una actitud positiva hacia las escuelas. Naturalmente, una pregunta no será suficiente; deben utilizarse preguntas relacionadas. Resulta bastante fácil obtener una definición de "nivel educativo". Se decide utilizar tres niveles: 1) licenciatura

FIGURA 25.1

	Actitud positiva	Actitud negativa
Licenciatura inconclusa		
Preparatoria terminada		
Preparatoria inconclusa		

inconclusa, 2) preparatoria terminada y 3) preparatoria inconclusa. El paradigma del análisis podría verse como el de la figura 25.1.

La virtud de paradigmas como éste es que el investigador puede determinar de inmediato si el problema específico se ha planteado con claridad y si está relacionado con el problema general. También proporciona cierta noción sobre cuántos encuestados se necesitarán para llenar adecuadamente las casillas de la tabla; también ofrece lineamientos para la codificación y el análisis. Además, como Katz (1953, pp. 80-81) afirma:

> Al realizar la tarea mecánica de diseñar dichas tablas, los investigadores seguramente descubrirán las complejidades de una variable que necesita medidas y calificaciones más detalladas de las hipótesis, en relación con las condiciones especiales.

El próximo paso en plan de flujo es la muestra y el plan del muestreo. El muestreo es demasiado complejo para discutirlo aquí en detalle, por lo que sólo se bosquejan las principales ideas. Para un tratamiento más detallado del tema, véase los capítulos 8 y 12 en esta obra, y el capítulo 5 del libro de Warwick y Lininger (1975). El ejemplo detallado del muestreo multietapas de área de Warwick y Lininger resulta especialmente útil. El *muestreo por área* es el tipo más utilizado en la investigación por encuesta. Primero deben definirse aleatoriamente las áreas grandes que van a muestrearse. Esto es equivalente a la partición del universo y al muestreo aleatorio de las casillas de la partición, las cuales pueden ser áreas delineadas por la cuadrícula de mapas o por fotografías aéreas de países, distritos escolares o manzanas de ciudades. Después se pueden obtener muestras de subáreas, a partir de las áreas grandes ya obtenidas. Finalmente, se eligen todos los individuos o familias o muestras aleatorias de individuos y familias.

En primer lugar, debe definirse el universo que va a estudiarse y del cual se obtendrán las muestras. ¿Se incluyen a todos los ciudadanos que viven en la comunidad? ¿A los líderes comunitarios? ¿A los ciudadanos que pagan impuestos escolares? ¿A quienes tienen hijos en edad escolar? Una vez definido el universo, se decide cómo se obtendrá la muestra y cuántos casos se elegirán. En la mejor investigación por encuesta se utilizan muestras aleatorias. Algunas ocasiones se utilizan las muestras *por cuota* en lugar de muestras aleatorias, debido a que estas últimas son costosas y más difíciles de llevar a cabo. En una muestra por cuota (o control por cuota), presumiblemente se logra la "representatividad" por medio de la asignación por cuotas a los entrevistadores —tantos hombres y mujeres, tantos blancos y afroamericanos, etcétera—. Aunque el muestreo por cuota puede lograr representatividad, carece de las virtudes del muestreo aleatorio —y, por lo tanto, debe evitarse—.

El siguiente gran paso en una encuesta es la construcción del inventario de entrevista y de otros instrumentos de medición que serán utilizados. Ésta es una tarea laboriosa y difícil que no tiene ninguna semejanza con los cuestionarios armados en forma apresurada en forma por los principiantes. La tarea principal consiste en traducir la pregunta de investigación en un instrumento de medición y en otros instrumentos construidos para la encuesta. Uno de los problemas del estudio podría ser, por ejemplo: ¿qué relación tienen las actitudes permisivas y restrictivas hacia la disciplina de los niños con los sistemas educativos locales? De entre las posibles preguntas para evaluar las actitudes permisivas y restrictivas, una sería: ¿cómo considera que debe disciplinarse a los niños? Después de completar bocetos de inventarios de entrevistas y de otros instrumentos, se prueban con antelación en una muestra pequeña representativa del universo. Después se revisan y se les da la forma final.

Los pasos descritos anteriormente constituyen la primera gran parte de cualquier encuesta. Después de que el investigador ha desarrollado el instrumento y ha determinado a qué población va a medir, también necesita decidir si los datos se recolectarán con el uso de un diseño transversal o con un diseño longitudinal. El diseño longitudinal implica la

Verificación de los datos obtenidos mediante encuestas

La investigación por encuesta posee una ventaja única entre los métodos científicos en ciencias sociales: con frecuencia es posible verificar la validez de los datos de la encuesta. Puede entrevistarse de nuevo a alguno de los entrevistados y comparar los resultados de ambos cuestionarios. Se ha encontrado que la confiabilidad de los reactivos de aspectos personales, como edad e ingreso, es alta. La confiabilidad de las respuestas de actitud es más difícil de determinar debido a que una respuesta modificada puede significar una actitud modificada. La confiabilidad de las respuestas promedio es más alta que la confiabilidad de las respuestas individuales. Por fortuna, el investigador generalmente está más interesado en promedios o medidas de grupo, que en respuestas individuales.

Una forma para verificar la validez de un instrumento de medición consiste en utilizar un criterio externo. Los resultados se comparan con un criterio externo, presumiblemente válido. Por ejemplo, un entrevistado declara haber votado en la última elección de miembros del consejo escolar. Es posible descubrir si esto es verdad o no, verificando los archivos de registro y de votación. Por lo común no se verifica el comportamiento individual, puesto que es difícil obtener información acerca de individuos; aunque la información grupal está más disponible. Esta información se utiliza para probar, hasta cierto grado, la validez de la muestra de la encuesta y de las respuestas.

Un buen ejemplo de una verificación externa de datos de encuesta es el uso de información sobre el último censo. Esto es particularmente útil en encuestas a gran escala; aunque también ayuda con encuestas más pequeñas. Es posible comparar proporciones de hombres y mujeres, razas, niveles educativos, edades, etcétera, de la muestra, con el censo de cada país. Por ejemplo, en el estudio de Verba y Nie (1972), sobre participación política, los autores informan ciertas comparaciones de este tipo. Los estimados de su muestra son precisos: sólo uno de ellos, *edad* 20-34, se desvía de los estimados del censo en más de 2 por ciento, lo cual constituye evidencia sólida de lo adecuado de la muestra. Para asegurarse, su muestra fue grande (> 2 500), pero también se han encontrado muestras pequeñas que son bastante precisas. En un estudio en Detroit, realizado por la University of Michigan, en 1952, la muestra fue de sólo 735, pero los estimados de la muestra estuvieron cerca de los del censo de 1950. Campbell y Katona (1953) analizan los métodos de verificación de la validez y confiabilidad de las muestras. Warwick y Lininger (1975) presentan tablas de errores de muestreo, con una explicación respecto a su significado y su uso estadístico. Ahí se observa, por ejemplo, que los porcentajes reportados entre 20 y 80, de una muestra de 700, tienen un error estándar de 4. ¡Para reducir el error estándar a 2, se requiere de una muestra de 3 000!

Los investigadores del SIDA, el doctor Vickie Mays y la doctora Susan Cochran, de la University of California en Los Ángeles, tienen una forma ingeniosa para verificar algunas de las respuestas en sus encuestas. Ellos incluyen en los cuestionarios reactivos que son específicos para ciertos grupos de personas. Por ejemplo, incluyen algunas preguntas que sólo incumben a hombres homosexuales, y otras preguntas dirigidas sólo a hombres heterosexuales. Posteriormente, después de que los datos se recolectan, se codifican y se traducen, de forma que la computadora pueda leerlos, se utiliza un programa estadístico de computadora para realizar una serie de tablas cruzadas de las preguntas respecto a la variable *preferencia sexual*. Los homosexuales que respondieron a las preguntas que eran sólo para hombres heterosexuales, o los heterosexuales que respondieron las preguntas para los homosexuales, se consideran como datos mal codificados, o capturados de manera incorrecta en la computadora. Entonces la forma de respuesta real puede recuperarse de los archivos y volverse a analizar para corregir los datos. Puesto que la realización de una encuesta significativa es costosa, cada cuestionario de cada participante es importante. A

diferencia de otras áreas de investigación, donde se recomendaría eliminar datos de participantes, la investigación por encuesta no puede hacer esto y aun así esperar obtener la información más precisa.

Tres estudios

Se han realizado muchas encuestas, tanto malas como buenas. Los estudiantes probablemente no se interesarían en la mayoría de ellas, ya que son sólo un poco más que intentos refinados para obtener información simple: estudios sobre la elección presidencial, sobre plantas industriales, etcétera. Sin embargo, existen encuestas de considerable interés y significancia —incluso muy grande— para los científicos del comportamiento. Tres de estos estudios se resumen a continuación.

Verba y Nie: participación política en Estados Unidos

Verba y Nie (1972) se preguntaron, entre otras cuestiones, cómo la participación política de los ciudadanos en una democracia influye en los procesos gubernamentales. Ellos entrevistaron a más de 2 500 residentes en Estados Unidos, en 200 lugares en 1967, los cuales fueron seleccionados por medio de un procedimiento de muestreo de probabilidad por área. (Sus comparaciones muestra-censo revelaron un alto acuerdo en general.) El principal hallazgo fue que la participación de los ciudadanos influye, de hecho, en los líderes políticos; pero son los ciudadanos más acaudalados, los más educados y con un estatus más alto en general, quienes más influyen con su participación. Los autores señalan que aunque los estadounidenses no se caracterizan por su ideología basada en la clase, el nivel social sí se relaciona con la participación. El estudio se caracteriza especialmente por su medición y metodología analítica sofisticadas, y por un hallazgo desconcertante importante. En capítulos posteriores se retomará el estudio y su metodología.

Docter y Prince: una encuesta de travestis masculinos

Docter y Prince (1997) reportan que una de las últimas encuestas más importantes publicadas sobre hombres travestis se llevó a cabo en 1972. El estudio de Docter y Prince (1997) utilizó el mismo instrumento de encuesta, usado en 1972, para medir a los travestis en 1992. Los investigadores añadieron algunas preguntas adicionales respecto al travestismo y a la excitación sexual. Uno de los objetivos consistía en evaluar si habían ocurrido cambios desde la encuesta de 1972. La razón por la que estos investigadores consideraron que quizá ocurrieron cambios, se centra en la discriminación del travestismo en algunas áreas de Estados Unidos, en la mayor exposición que han tenido los travestis y los transexuales en los medios de comunicación masiva y en el crecimiento de grupos y organizaciones nacionales de apoyo. Docter y Prince compararon las dos muestras encuestadas en, por lo menos, seis dimensiones: 1) factores demográficos, de la infancia y familiares; 2) orientación sexual y comportamiento sexual; 3) identidad de género; 4) comportamiento del papel de género; 5) planes futuros de vivir completamente como mujer, y 6) confianza en los servicios de asesoría o de salud mental. Docter y Prince utilizan el término "travesti" para definir a los hombres biológicos que ocasionalmente se visten con ropa de mujer, que, sin embargo, no buscan una reasignación de sexo. Un transgenérico es aquel que vive continuamente en el papel del género opuesto a su sexo biológico, sin procedimientos de reasignación de sexo; un transexual es quien ya tuvo una reasignación de sexo. Docter y Prince encuestaron a 1 032 travestis autodefinidos, en edades de 20 a 80 años. Biológicamente todos eran

hombres. La población de la muestra fueron voluntarios de todo Estados Unidos, quienes respondieron a la solicitud de participantes para investigación en reuniones de clubes, convenciones y publicaciones para travestis. La muestra de 1992 representó una base más amplia de travestis que la muestra de 1972: la de 1972 consistió principalmente de lectores de publicaciones para travestis; en cambio, la muestra de 1992 estuvo compuesta de lectores de un número de diferentes publicaciones y de miembros de clubes de travestis. La comparación entre las dos muestras demostró que hubo algunos cambios entre los travestis de 1972 y los de 1992. En particular, en la muestra de 1992 existían más participantes interesados en vivir como mujeres todo el tiempo. También hubo más participantes en la muestra de 1992 que tenían una identidad de género preferida, que fue igual para hombre que para mujer, con respecto a la muestra de 1972. Docter y Prince (1997) documentan las diferencias en el método de muestreo entre las dos muestras y señalan las limitaciones de la muestra más reciente, en comparación con la muestra antigua. Ésta es la naturaleza de la investigación por encuesta: ciertas cuestiones cambian a través del tiempo, y vuelven más difícil obtener exactamente el mismo ambiente de investigación, de un periodo a otro.

Sue, Fujino, Hu, Takeuchi y Zane: servicios comunitarios de salud para minorías étnicas

Este estudio (1991) no se ajusta exactamente a lo que se llamaría investigación por encuesta. Tales investigadores no diseñaron la encuesta para el estudio, ni recolectaron los datos para el mismo. En su lugar, utilizaron los datos suministrados por el Automated Information System (AIS), mantenido por el Departamento de Salud Mental del condado de Los Ángeles. Estos datos fueron utilizados por la agencia del gobierno con el propósito de la administración del sistema, recaudación fiscal, manejo clínico e investigación. Todos los pacientes eran receptores de consulta externa. El estudio califica como investigación por encuesta debido a que se trata de un estudio de campo —cuantitativo y epidemiológico— que reunió información que describe las relaciones entre variables dentro del conjunto de datos. Dicho tipo de investigación por encuesta está basado en la búsqueda de registros (Isaac y Michael, 1987). Sue y sus colaboradores utilizaron los datos para responder algunas preguntas respecto a los servicios de salud mental de cuatro grupos étnicos: afroamericanos, asiaticoamericanos, latinos y estadounidenses blancos. La muestra del AIS consistió de 7 136 asiaticoamericanos, 47 220 afroamericanos, 58 844 latinos y 99 036 estadounidenses blancos. El conjunto de datos originales cubrió un periodo de 15 años. Sue y sus colaboradores utilizaron únicamente el último periodo de cinco años. En una comunicación personal, Sue informó al segundo autor de este libro (HBL) que él y su equipo dedicaron gran cantidad de tiempo, esfuerzo y dinero en la reorganización de los datos, para que pudieran ser sujetos de su investigación. La hipótesis que los investigadores comprobaron fue que los pacientes que fueron apareados tanto étnicamente como por género con un terapeuta, tendrían una mayor mejoría en su salud mental. La medida de salud mental fue la escala de evaluación global (EEG). Las variables dependientes en dicho estudio fueron los abandonos del tratamiento, el número promedio de sesiones de tratamiento y los resultados del tratamiento. Los resultados del estudio mostraron en todos los grupos, con excepción de los afroamericanos, menores posibilidades de abandono del tratamiento cuando los pacientes estaban apareados étnicamente con el terapeuta. Cuando se aparearon por género, tan sólo los asiaticoamericanos y los estadounidenses blancos demostraron una menor posibilidad de abandono del tratamiento. Tal hallazgo señala un ingrediente importante para la prevención del abandono de los pacientes, en instituciones públicas de salud mental. Al analizar los datos del AIS, Sue y sus colaboradores encontraron, que sólo una tercera parte de los pacientes étnicos fueron tratados por terapeutas de la misma etnia; mientras que el 75 por ciento de los estadounidenses blancos fueron tratados por terapeutas blancos. Respecto al número promedio de sesiones de tratamiento, todos los grupos que

incluían un apareamiento paciente-terapeuta tuvieron un número promedio mayor de sesiones de tratamiento. No obstante, respecto a la variable de apareamiento por género, únicamente los mexicoamericanos y los estadounidenses blancos mostraron un número mayor de sesiones de tratamiento. El EEG fue utilizado para medir el resultado del tratamiento. No hubo un efecto del apareamiento por género. Respecto al apareamiento étnico, sólo los mexicoamericanos tuvieron puntuaciones más altas en el EEG en el momento de la conclusión del tratamiento.

Hall, Kaplan y Lee (1994) hallaron patrones similares, al utilizar la misma base de datos, pero observando únicamente a los pacientes que eran niños. Encontraron que los niños más pequeños tenían una mayor mejoría cuando eran apareados con terapeutas similares en las áreas de etnicidad e idioma. Esto puede atribuirse al hecho de que el idioma, en niños bilingües más pequeños, aún no está bien desarrollado, lo cual resulta en la necesidad de un terapeuta que pueda satisfacer sus requerimientos culturales y de idioma. Otro estudio originado a partir de esta importante base de datos es el de Russell, Fujino, Sue, Cheung y Snowden (1996).

Aplicaciones de la investigación por encuesta en educación

Estos estudios demuestran con claridad la aplicabilidad de la investigación por encuesta y su metodología en sociología, psicología social, trabajo social, psicología clínica y ciencias políticas. El fuerte énfasis de la investigación por encuesta en las muestras representativas, el diseño general, el plan de investigación y las entrevistas a expertos con el uso de inventarios de encuesta construidos cuidadosa y competentemente, ha sido, y continuará siendo, una influencia benéfica en la investigación del comportamiento. A pesar de su evidente valor potencial en todos los campos de la investigación del comportamiento, la investigación por encuesta no ha sido utilizada en tan amplia medida en donde parecería tener un gran valor teórico y práctico, es decir, en la educación. Su utilidad distintiva en la educación y en la investigación educativa parece haber sido lentamente percibida. No obstante, una revisión de la literatura actual muestra que la situación quizás esté cambiando. Por lo tanto, se dedica esta sección a la aplicación de la investigación por encuesta a la educación y a los problemas educativos.

Obviamente, la investigación por encuesta constituye una herramienta útil para indagar hechos en la educación. Un administrador, un consejo educativo o un grupo de maestros aprenden mucho acerca de un sistema escolar o de una comunidad sin contactar a cada niño, cada maestro ni cada ciudadano. En resumen, los métodos de muestreo desarrollados en la investigación por encuesta resultan de mucha utilidad. Es poco satisfactorio depender de las llamadas muestras representativas que son relativamente de ensayo y error, basadas en juicios "expertos". Tampoco es necesario reunir datos en poblaciones completas; las muestras son suficientes para muchos propósitos.

La mayor parte de la investigación en educación se conduce utilizando muestras no aleatorias, relativamente pequeñas. Si las hipótesis se sostienen, pueden comprobarse después con muestras aleatorias de poblaciones y, si nuevamente son apoyadas, entonces pueden generalizarse los resultados a poblaciones de escuelas, niños y gente ordinaria. En otras palabras, la investigación por encuesta puede usarse para probar hipótesis que ya han sido probadas en situaciones más limitadas, lo cual resulta en un incremento de la validez externa.

La investigación por encuesta parece ajustarse de manera ideal a algunos de los aspectos controvertidos más importantes en educación. Por ejemplo, su habilidad para manejar problemas "difíciles" como la integración y la clausura de escuelas, a través de entrevistas

cuidadosas y prudentes, la ubica en los primeros lugares de la lista de los métodos de investigación para dichos problemas. Las entrevistas de muestras aleatorias de ciudadanos y maestros de distritos escolares, al inicio de un programa de educación especial o de educación para superdotados, o tras la experiencia de la clausura probable de ciertas escuelas primarias a causa de la disminución de inscripciones, pueden ofrecer información valiosa respecto a sus preocupaciones y temores, de tal forma que se tomen medidas apropiadas para informarles y para disminuir sus temores. El efecto de estas medidas, por supuesto, se estudia también.

La investigación por encuesta quizá se adapta más a la obtención de hechos personales y sociales, creencias y actitudes. Es significativo que, aunque se dicen y escriben cientos de miles de palabras sobre educación y sobre lo que se supone que la gente piensa acerca de ella, existe poca información confiable sobre el tema. Simplemente no se conocen las actitudes que tiene la gente hacia la educación. Es necesario depender de futuros escritores y de los llamados expertos, para obtener dicha información. Los consejos educativos con frecuencia dependen de administradores y de líderes locales para que les digan lo que la gente piensa. Algunas de las preguntas que se plantean y, que, posiblemente, se contestan con el uso de la investigación por encuesta son: ¿apoyará la comunidad un mayor presupuesto el próximo año? ¿Qué pensarán respecto a la división de los distritos escolares? ¿Cómo reaccionarán los padres ante la asignación de tareas a los niños para lograr suprimir la segregación racial? ¿Cuál es el currículum actual? ¿Cuál es la tasa de abandono de los estudiantes de posgrado? ¿En qué grado copian más en los exámenes los estudiantes de la escuela de medicina? ¿Los niños con diversos antecedentes culturales que viven en Israel difieren respecto a sus temores? Un antiguo y sobresaliente ejemplo de investigación por encuesta en educación es el estudio de Gross, Mason y McEachern (1958), el cual constituye una lectura indispensable para los administradores educativos y para los miembros de los consejos educativos.

Es alentador que en los pasados 12 años se hayan realizado más estudios en ambientes educativos. Considere, por ejemplo, el estudio de Stile, Kitano, Kelley y Lecrone (1993), quienes llevaron a cabo una encuesta nacional sobre lo que está sucediendo en programas preescolares y de jardín de niños para alumnos superdotados. Su encuesta examinó escuelas en todas las entidades, en Estados Unidos. Reportaron que tan sólo 29 de los 50 estados (58 por ciento) y un territorio tenían programas para niños superdotados. Dichos programas abarcaron un poco más de 2 655 distritos escolares. Sólo 16 estados muestran que tienen programas en el nivel de jardín de niños para aprendices superdotados, que provienen de familias en desventaja. Aunque el estudio de Stile y colaboradores parece, de alguna manera, una encuesta sobre estatus, sí señala cómo se ve "el panorama completo" de los programas educativos para superdotados en Estados Unidos y sus territorios. También señala el tipo de fondos utilizados para los programas con superdotados.

El estudio de Cooke, Sims y Peyrefitte (1995) brinda información que no se había publicado antes, respecto al abandono escolar de estudiantes de posgrado. Se conoce mucho acerca del abandono escolar de estudiantes de licenciatura; pero muy poco sobre los estudiantes de posgrado. Por lo común, el muestreo de estudiantes de posgrado no es tan abundante como el de los estudiantes de licenciatura. En dicho estudio los investigadores reunieron datos de 230 estudiantes de posgrado inscritos en programas de negocios, ingeniería, administración pública y educación. Se eligieron estos programas debido a que se contaba con mayores cantidades de estudiantes de minorías étnicas inscritos en ellos. El instrumento de encuesta se envió por correo a los participantes a principios de 1992; una encuesta de seguimiento se envió 18 meses después. Las dos encuestas fueron utilizadas para determinar si podía predecirse el abandono después de 18 meses. Los resultados demostraron que las minorías étnicas tenían una mayor intención de renunciar a los estu-

dios de posgrado, y que estaban menos satisfechos con los estudios de posgrado que aquellos que no pertenecían a tales minorías étnicas. Sin embargo, aunque estas diferencias existían, no se encontró que estuvieran relacionadas con el abandono escolar. El abandono estuvo más relacionado con las variables necesidad de logro, compromiso afectivo y si el programa de posgrado cumplía las propias expectativas.

Little y Lee (1995) realizaron un estudio sobre todos los programas escolares de posgrado en psicología, a través de todo el territorio de Estados Unidos. Su propósito era determinar la cantidad de entrenamiento que recibían los estudiantes de posgrado, en las áreas de métodos de investigación y estadística. Entre las muchas comparaciones, está la que se hizo entre los programas que otorgaban doctorados y los que no lo hacían. Little y Lee no estaban interesados únicamente en la cantidad de cursos, sino también en su contenido y en el empleo de programas estadísticos de computadora. Se enviaron por correo un total de 181 encuestas a los programas certificados por la National Association of School Psychologists (NASP) y a la American Psychological Association (APA), así como a aquellos anotados en la *Petersen's Guide to Graduate Education*. De éstos, se obtuvieron 101 encuestas útiles. Los resultados no mostraron diferencias significativas dentro de los programas predoctorales y doctorales, respecto a la cantidad de cursos sobre estadística y diseño de investigación. Sin embargo, se encontraron diferencias al comparar los programas predoctorales con los de doctorado, los cuales, por lo general, requerían el doble de cursos de estadística y de diseño de investigación, que los programas predoctorales. Little y Lee proporcionan información valiosa que puede utilizarse en programas de posgrado ya existentes, o nuevos, en psicología escolar, para ajustar o desarrollar su currículum.

Baldwin, Daugherty, Rowley y Schwarz (1996) enviaron una encuesta a 3 975 estudiantes de segundo año que acudían a 31 escuelas de medicina; 2 459 (62 por ciento) completaron el cuestionario de encuesta. La encuesta se llevó a cabo para determinar el grado de comportamiento y actitud fraudulentos en los exámenes. El 39 por ciento de los encuestados afirmaron haber visto por lo menos un incidente en donde se hiciera trampa en un examen. Cerca de dos terceras partes de la muestra declaran haber escuchado que los estudiantes cometen fraude en los exámenes. El hecho fraudulento se dividió en categorías: 1) obtener información previa respecto al examen, 2) copiar las respuestas de otro estudiante durante el examen y 3) intercambiar respuestas durante el examen. El 82 por ciento de los estudiantes que declararon haber hecho trampa en la escuela de medicina también afirmaron haberla hecho en la escuela, antes de ingresar a la facultad de medicina. Aproximadamente el 5 por ciento de los estudiantes reportaron haber hecho trampa alguna vez durante los primeros dos años de la escuela de medicina. Más hombres que mujeres afirmaron haber realizado dichas prácticas en los exámenes.

Ventajas y desventajas de la investigación por encuesta

La investigación por encuesta tiene la ventaja de una visión amplia: se obtiene gran cantidad de información a partir de una población grande. Permite estudiar una población grande o un sistema escolar grande, a mucho menor costo que el que generaría un censo. Mientras que las encuestas tienden a ser más costosas que los experimentos de laboratorio y de campo, y aun que los estudios de campo, por la cantidad y calidad de la información que brindan, resultan económicas. Además, es posible utilizar las instalaciones y el personal educativo para reducir los costos de la investigación.

La información de la investigación por encuesta es precisa —con cierto error de muestreo, por supuesto—. La precisión de las muestras obtenidas apropiadamente es, con

frecuencia, sorprendente incluso para los expertos en el campo. Una muestra de 600 a 700 individuos o familias brinda un retrato notablemente preciso de una comunidad: sus valores, actitudes y creencias.

Aunadas a estas ventajas, están las debilidades y desventajas inevitables. Una primera desventaja es que la información de las encuestas generalmente no profundiza mucho debajo de la superficie. Por lo común, la visión de la información buscada se enfatiza a expensas de la profundidad. No obstante, ésta parece constituir una debilidad que no es necesariamente inherente al método. Los estudios de Verba y Nie (1972), y de Smith y Garner (1976; véase también Garner y Smith, 1977) muestran que es posible ir considerablemente a mayor profundidad de las opiniones superficiales. Smith y Garner diseñaron un procedimiento para acompañar un cuestionario bien diseñado, que les permitió penetrar en el comportamiento homosexual de atletas universitarios. En lugar de aplicar un instrumento de encuesta una vez, ellos lo aplicaron por lo menos tres veces para verificar la consistencia de las respuestas. También desarrollaron otros medios de verificación de las respuestas de los atletas, respecto a un tema muy sensible y emplearon medios totalmente inofensivos para recolectar sus datos. Smith y Garner obtuvieron información útil respecto a un tema altamente emocional. A pesar de estos ejemplos sobre la profundidad de la información de la investigación por encuesta, ésta parece adaptarse mejor a la investigación extensiva, que a la intensiva. Otros tipos de investigación quizás están mejor adaptados para una exploración más profunda de algunas relaciones.

La segunda desventaja es de tipo práctico. La investigación por encuesta demanda mucho tiempo, energía y dinero. En una encuesta grande, pueden pasar varios meses antes de que una sola hipótesis llegue a probarse. El muestreo y el desarrollo de buenos inventarios son operaciones importantes. Las entrevistas requieren habilidad, tiempo y dinero. Las encuestas a menor escala pueden evitar dichos problemas en cierto grado.

Cualquier investigación que utilice muestreo está sujeta, naturalmente, al error de muestreo. Aunque es verdad que la información de las encuestas ha mostrado ser relativamente precisa, siempre existe una probabilidad de 20 o 100 de que ocurra un error —más serio de lo que podrían causar fluctuaciones pequeñas por azar. La probabilidad de dicho error puede disminuirse al crear verificaciones de seguridad en un estudio— si se incluyen comparaciones con datos de censos u otra información externa, y por medio del muestreo independiente de la misma población.

Una debilidad potencial, más que real, de este método es que la entrevista de encuesta saca temporalmente al entrevistado de su propio contexto social, lo que puede invalidar los resultados de la encuesta. La entrevista es un suceso especial en la vida ordinaria del entrevistado. Tal vez esa diferencia haga que el entrevistado hable e interactúe con el entrevistador de una manera poco natural. Por ejemplo, una madre, al ser interrogada acerca de las prácticas del cuidado de sus hijos, quizá dé respuestas que revelen métodos que a ella le gustaría utilizar, en lugar de los que en realidad utiliza. Los entrevistadores tienen la oportunidad de limitar los efectos de sacar a los entrevistados de su contexto social, por medio de un manejo experto, en especial con una manera propia y por medio de expresar y plantear las preguntas de manera cuidadosa (véase Cannell y Kahn, 1968).

La investigación por encuesta también requiere de bastante conocimiento y sofisticación sobre investigación. El investigador por encuesta competente debe saber sobre muestreo, sobre la construcción de preguntas e inventarios, la forma de realizar entrevistas, el análisis de datos y otros aspectos técnicos de la entrevista. Dicho conocimiento es difícil de adquirir; pocos investigadores alcanzan este tipo y cantidad de experiencia. Conforme se va apreciando el valor de la investigación por encuesta, tanto de gran escala como de pequeña escala, puede anticiparse que dicho conocimiento y experiencia serán considerados, por lo menos de forma mínima, necesarios para los investigadores.

Meta-análisis

En el momento de escribir estas líneas va en aumento el número de estudios de investigación reportados que utilizan el meta-análisis. El estudiante que lea atentamente la literatura, tiene probabilidad de encontrarse con un estudio que utilice el meta-análisis. Pero, ¿qué es el *meta-análisis*, y por qué se incluye dentro de la investigación por encuesta? Bueno, muchos escritores de libros de texto han tenido problemas para ubicar estos métodos dentro de capítulos de temas específicos. Robert Rosenthal, una de las principales autoridades sobre dicho tema, ubicó el tema del meta-análisis en el apéndice del libro que escribió junto con Ralph Rosnow sobre investigación del comportamiento (Rosnow y Rosenthal, 1996). Algunos autores lo han integrado dentro de capítulos que tratan sobre estadística. Aquí no es diferente. Los autores perciben este método como perteneciente a la investigación por encuesta. Aunque no se diseña ningún cuestionario ni se planea ninguna muestra, sí implica la búsqueda de datos previamente recolectados. Tales datos provienen de la literatura de investigación. Podría decirse que es un tipo de encuesta de la literatura. El meta-análisis es de naturaleza cuantitativa y no experimental. Algunos autores, como Mann (1990), se han referido a éste como un experimento natural. El propósito del meta-análisis es buscar en la literatura un tema específico que contenga un gran número de estudios. Algunos de dichos estudios pueden coincidir entre sí de alguna manera. Si es así, producen una convergencia de conocimiento, y ese conocimiento se vuelve útil en la toma de decisiones. Por ejemplo, si existe un efecto de la preparación para el Scholastic Aptitud Test (SAT), entonces todos los estudios realizados sobre el tema deben tener hallazgos básicos similares. El meta-análisis implica tomar todos estos estudios de forma colectiva, para determinar si un hallazgo similar se encuentra una y otra vez bajo situaciones diferentes. La meta es ser capaz de establecer algún tipo de ley general del comportamiento. A diferencia de los estudios de investigación "regulares", que tienen a los participantes individuales o grupos de participantes como unidad de medición, el meta-análisis utiliza los estudios individuales como unidades de medición. Los resultados de estos estudios de investigación son resumidos por medio del uso de medidas del tamaño del efecto, similares a la ETA cuadrada (η^2) o a la omega cuadrada (ω^2), que se analizaron en un capítulo previo y que se utilizan para estudios de investigación individual. En el meta-análisis, el tamaño del efecto se mide utilizando un estadístico *d*. Gran parte de la investigación meta-analítica que se utiliza hoy, reporta de manera tabulada los diferentes estudios, el tamaño de la muestra y el tamaño del efecto. La tabla 25.2 presenta una adaptación del estudio de Scogin y McElreath (1994) sobre la efectividad de la intervención psicológica en la depresión de adultos mayores.

Para determinar el efecto general del fenómeno bajo estudio, Rosenthal (1978) ofrece un procedimiento estadístico para calcular el tamaño del efecto combinado. Rosenthal

TABLA 25.2 *Tabla meta-analítica del tamaño de la muestra y del tamaño del efecto (de Scogin y McElreath)*

Estudio	Tamaño de la muestra	Tamaño del efecto
1	31	.41
2	36	.00
3	84	.97
4	61	.70
5	28	.82
6	20	.28

toma, en esencia, los valores de p[1] de cada estudio, encuentra la puntuación estándar para cada valor de p, y la utiliza en la fórmula:

$$Z_{general} = \frac{\sum_{i=1}^{n} Z_i}{\sqrt{n}}$$

Después se determina la probabilidad de este valor Z, a través del uso de la tabla de la distribución normal (véase apéndice B). Esto indicará al investigador si el efecto combinado general de los estudios es estadísticamente significativo o no. Por lo tanto, en un meta-análisis el investigador puede encontrar una gran cantidad de estudios respecto a un fenómeno particular, que no fueron estadísticamente significativos. Sin embargo, al combinarse, es posible lograr la significancia estadística; por ejemplo, un investigador encuentra cuatro estudios que tienen los siguientes valores p: .25, .32, .04, .19, para pruebas de una cola. Sus valores Z correspondientes son .69, .47, 1.75 y .50, respectivamente. Observe que sólo el último de estos estudios fue estadísticamente significativo. El valor Z general para este ejemplo sería:

$$Z_{general} = \frac{0.69+0.47+1.75+0.50}{\sqrt{4}} = \frac{3.41}{2} = 1.72$$

El valor Z general es significativo al nivel 0.0427. Rosenthal (1978) muestra nueve formas de reunir resultados de estudios para crear un estadístico general.

El meta-análisis no debe confundirse con otros dos métodos similares: la réplica y el análisis con diferentes modelos o métodos. En la réplica se utilizan la misma metodología y los mismos datos recolectados, de una muestra diferente. La meta de los estudios de réplica es establecer la confiabilidad de los resultados en la misma situación. En el modelo del uso de diferentes métodos, se recolectan los mismos datos de una muestra diferente o se utilizan los datos originales; sin embargo, en este método se utilizan diferentes métodos. El objetivo aquí es indagar qué tan robustos fueron los hallazgos originales. Se hace un esfuerzo por encontrar el modelo que mejor ajuste para realizar predicciones o para tomar decisiones. Ésta puede ser considerada una forma de "extracción de datos", o investigación de métodos múltiples. El meta-análisis esencialmente combina estos dos, la búsqueda de métodos diferentes y de datos diferentes. El objetivo del meta-análisis es generalizar los resultados a situaciones nuevas.

El desarrollo del meta-análisis se acredita a Glass (1976). Smith y Glass (1977) demostraron el meta-análisis en una búsqueda a través de la literatura psicológica, realizada para determinar la eficacia de la psicoterapia. Encontraron cerca de 400 estudios que ofrecían información sobre psicoterapia relevante para su meta. Smith y Glass fueron capaces de sintetizar los resultados de cada uno de estos estudios, para establecer una conclusión general acerca de la eficacia de la psicoterapia. Además, pudieron comparar la eficacia relativa de varios métodos distintos de tratamiento dentro de la psicoterapia. Por lo tanto, el meta-análisis es capaz de responder gran cantidad de preguntas prácticas y de investigación que la investigación individual no puede lograr. Por ejemplo, Blumenthal (1998)

[1] El valor p es otra forma de expresión de la probabilidad de un error tipo I. Generalmente, los estudios reportan los resultados como $p < .05$ (estadísticamente significativa) o $p > .05$ (estadísticamente no significativa). No obstante, en años recientes, con la disponibilidad de las computadoras de alta velocidad y de los programas de computadora, el valor p se calcula directamente.

utiliza un meta-análisis para responder preguntas respecto a las diferencias de género en la percepción del acoso sexual. Este estudio es muy significativo en el área de casos legales y de la corte, o en políticas legales. La mayor parte de los estudios sobre el acoso sexual han incluido participantes a quienes se les presenta una o dos escenas breves sobre un incidente, y después se les plantea una serie de preguntas respecto a la situación. Mientras que la mayoría de los estudios presentan resultados en donde los hombres y las mujeres difieren en su percepción del acoso sexual, la magnitud de los hallazgos ha variado. Inclusive existen algunos estudios que no encontraron diferencias significativas. El estudio de Blumenthal examina la literatura sobre este tema y determina, de manera sistemática, cuál es la situación general en tal aspecto. El estudio de Blumenthal utiliza búsquedas por computadora de estudios con palabras clave tales como "acoso sexual", "percepción" y "diferencias por género". El advenimiento de las búsquedas por computadora ha facilitado el crecimiento de los estudios meta-analíticos.

Antes del desarrollo del meta-análisis, los investigadores se basaban en artículos que aparecían en publicaciones especializadas de revisión, tales como *Psychological Review*, *American Psychologist*, *Psychological Bulletin*, *Harvard Education Review* y el *Annual Review of Psychology*, para encontrar resúmenes de investigación que hubiesen sido realizados en una cierta área. Los escritores de las revisiones eran elegidos, por lo general, debido a que se les consideraba expertos en esa área. A pesar de que los revisores hacían grandes esfuerzos para presentar los datos de forma objetiva, era inevitable cierto nivel de subjetividad. Mann (1990) presenta algunos ejemplos de revisiones subjetivas, realizadas con el modelo tradicional en la ciencia médica. Mann afirma que siempre existe la posibilidad de que ciertos elementos importantes puedan pasarse por alto al hacer una de estas revisiones tradicionales. El meta-análisis proporciona una metodología que suplementa estas revisiones y que satisface una necesidad crítica en la ciencia. Esa necesidad es la resolución de hallazgos de investigación conflictivos. Simon (1987) considera que el meta-análisis no podría resolver el conflicto por completo. Basa su argumento en la premisa de que los estudios meta-analíticos no toman en cuenta suficientes variables independientes. Si se considera el problema planteado por Adelson y Williams, reportado en Simon (1987), el meta-análisis no sería capaz de responder la pregunta sobre cuál de las 34 variables independientes posibles ejerce el mayor efecto sobre el desempeño del piloto.

Para ilustrar algunas de las áreas adecuadas para el meta-análisis, se citarán algunos de los estudios que se han realizado. Scogin y McElreath (1994) realizaron un meta-análisis de 17 estudios respecto a la eficacia de los tratamientos psicosociales para la depresión de adultos mayores. Estos 17 estudios cumplieron el criterio de los estudios que tuvieron una condición de control. Los investigadores buscaron en la literatura artículos relacionados con el tema, publicados entre 1975 y 1990. El promedio del tamaño del efecto encontrado por estos investigadores, resultó estadísticamente significativo e indicó que aquellos que recibieron tratamiento psicosocial estaban más saludables que quienes no recibieron dicho tratamiento. Verhaeghan y DeMeersman (1998) realizaron un meta-análisis de estudios que comparaban adultos mayores y adultos jóvenes sobre el efecto de interferencia de Stroop. El efecto de Stroop ya fue mencionado en un capítulo anterior. Los participantes del estudio nombraban el color de la tinta con que estaba escrita la palabra, en lugar de la palabra misma. Es decir, con la palabra *amarillo* escrita con tinta color verde, ellos tendían a decir "verde", cuando se les pedía que leyeran la palabra. Verhaeghan y DeMeersman encontraron 20 estudios en una búsqueda de literatura por computadora. Los hallazgos de Verhaeghan y DeMeersman demostraron que el efecto de Stroop no se vio afectado por la edad. Hellman (1997) utilizó un meta-análisis para estudiar la relación entre la satisfacción laboral y la intención de dejar el trabajo. Hellman identificó 38 estudios en la búsqueda de la literatura en un periodo de 13 años. La relación entre estas dos variables se encontró,

de forma consistente estadísticamente significativa y en dirección negativa. Es decir, a mayor satisfacción, menor posibilidad de dejar el trabajo; o a menor satisfacción, mayores intentos por dejarlo.

El meta-análisis es un método que puede resumir los resultados de muchos estudios realizados respecto a la misma o similar área temática. No requiere que los estudios estén replicados de manera exacta. Además, posee el apoyo de por lo menos un índice cuantitativo —tamaño promedio del efecto— como ayuda en la evaluación. Además, los índices del tamaño del efecto también pueden compararse estadísticamente entre sí. Sin embargo, existe por lo menos un problema que se ha asociado con el meta-análisis: el "problema del cajón de archivo", el cual surge del hecho de que los editores de las revistas generalmente no aceptan publicar artículos que tengan resultados "no significativos". Es decir, estudios donde no se rechaza la hipótesis nula o estudios que no sean estadísticamente significativos al místico nivel $\alpha = .05$. Barber (1976) lo llama el "efecto negativo". Dichos estudios son considerados "impublicables". Por lo tanto, el meta-análisis, que generalmente se realiza para revisar la literatura de investigación y encontrar artículos publicados del área de interés, contendrá sólo el análisis y las conclusiones extraídas a partir de estudios que sean "estadísticamente significativos". Mientras tanto, los investigadores pueden haber almacenado en sus "cajones de archivo" los estudios de investigación que fracasaron en la producción de resultados significativos. Si un cajón de archivo de este tipo existe, con un gran número de hallazgos no significativos, entonces los meta-análisis reportados por los investigadores pueden ser exageradamente optimistas. Para contrarrestar este problema, algunos investigadores han desarrollado tablas y métodos para dar al investigador alguna idea de la cantidad de tolerancia o distorsión que pudiera estar presente (véase Bradley y Gupta, 1997; Sharpe, 1997). Sin embargo, Rosenthal y Rubin (1978) han desarrollado una fórmula estadística para determinar cuántos estudios negativos se necesitarían para echar por tierra la conclusión extraída con el uso de estudios positivos en un meta-análisis. Rosenthal y Rubin han mostrado que por 345 estudios publicados, se necesitarían 65 123 estudios no publicados, que mostraran un efecto negativo. Sin embargo, como Light y Pillemer (1984) han señalado, existe una diferencia importante entre 50 estudios sin efecto no publicados, y 50 000 estudios sin efecto no publicados, aun cuando ambos estén por debajo de los 65 123 estipulados por Rosenthal y Rubin.

El análisis estadístico del meta-análisis puede ser bastante complejo. Se hizo una breve mención acerca del tamaño del efecto calculado. Sin embargo, el análisis va más lejos que éste. Existen numerosos programas de cómputo disponibles para manejar los cálculos (véase Johnson, 1993; Mullen, 1993).

RESUMEN DE CAPÍTULO

1. La investigación por encuesta es un tipo de estudio de campo cuantitativo.
2. La investigación por encuesta intenta encontrar relaciones entre variables sociológicas y psicológicas.
3. La investigación por encuesta es un desarrollo del siglo XX.
4. El enfoque general de la investigación por encuesta es la gente.
5. Las entrevistas, los inventarios, los paneles y las encuestas por teléfono y por correo constituyen diferentes tipos de encuestas.
6. El tipo de encuesta que produce la mejor información es la entrevista. Las encuestas por correo son las que contienen la mayor cantidad de problemas.
7. La investigación por encuesta puede obtener un amplio rango de información, pero no proporciona información profunda. Es más extensiva que intensiva.

8. La metodología de la investigación por encuesta incluye un "plan de flujo". Este plan bosqueja el diseño y la implementación de una encuesta.
9. La construcción del cuestionario o encuesta es una de las partes importantes del plan. Otra parte importante es el plan de muestreo (por ejemplo, ¿a quién se va a muestrear y cómo se llevará a cabo el muestreo?).
10. La recolección de datos es con frecuencia una tarea laboriosa. Si se utilizan entrevistas, entonces el entrevistador necesita ser entrenado apropiadamente.
11. Convertir los datos a una forma que la computadora pueda leer es otra gran tarea en la investigación por encuesta. Esto también incluye el análisis de los datos.
12. La investigación por encuesta puede ser costosa en términos del tiempo, dinero y trabajo. En una encuesta grande, los hallazgos no son accesibles rápidamente antes de finalizar el estudio.
13. El meta-análisis es una forma de investigación por encuesta. La investigación experimental generalmente utiliza un participante individual como unidad de medición. En el meta-análisis, los estudios individuales son, por sí mismos, la unidad de la medición.
14. El meta-análisis implica recolectar una cantidad de estudios sobre un tema similar y resumir los hallazgos. La meta es definir algunas leyes generales de comportamiento.

SUGERENCIAS DE ESTUDIO

1. A continuación se muestran varios buenos ejemplos de investigación por encuesta; algunos son artículos y otros son libros. Elija uno de ellos. Si elige un libro, lea el primer capítulo para aprender sobre el problema del estudio. Después vaya a la sección técnica (si existe una), para ver cómo se realizaron el muestreo y la entrevista. (La mayoría de los estudios de investigación por encuesta publicados contienen dicha sección.) Intente determinar las variables principales y sus relaciones. Dentro de los corchetes se incluyen resúmenes del contenido.

Cai, D. y You, M. (1998). An ergonomic approach to public squatting-type toilet design. *Applied Ergonomics, 29*, 147-153. [Un estudio sobre el diseño de un tipo de retrete público.]

Glock, C. y Stark, R. (1966). *Christian beliefs and anti-Semitism.* Nueva York: Harper y Row. [Religión y prejuicio.]

Lortie, D. (1975). *Schoolteacher: A sociological study.* Chicago: University of Chicago Press. [Un estudio valioso y revelador sobre los maestros.]

MacDonald, S. Wells, S. y Lothian, S. (1998). Comparison of lifestyle and substance use factors related to accidental injuries at work, home, and recreational events. *Accident Analysis and Prevention, 30*, 21-27.

Miller, W. y Levitin, T. (1976). *Leadership and change: The new politics and the American electorate.* Cambridge, Massachusetts.: Winthrop. [La "nueva izquierda" y la "minoría silenciosa". Con base en los datos de 25 años del Centro de Investigación por Encuesta.]

Murray, A. (1998). The home and school background of young drivers involved in traffic accidents. *Accident Analysis and Prevention, 30*, 169-182. [Investiga la relación entre los antecedentes del hogar y escolares, y los datos sobre accidentes de más de 4 000 conductores masculinos y femeninos, de edades entre 16 y 22 años.

Oates, G. L. (1997). Self-esteem enhancement through fertility? Socioeconomic prospects, gender, and mutual influence. *American Sociological Review, 62*, 965-973. [Determina si tener hijos influye o no en la autoestima de las personas.]

2. Rensis Likert fue un científico social sobresaliente, un pionero metodológico en la investigación por encuesta y fundador del Instituto de Investigación Social de la University of Michigan (de la cual forma parte el Centro de Investigación por Encuesta). Dos de sus colegas, Seashore y Katz (1982) escribieron un obituario en el que describieron las contribuciones de Likert. Se sugiere que los estudiantes lean el obituario, el cual es virtualmente una explicación del nacimiento y crecimiento de aspectos metodológicos importantes de la investigación por encuesta, así como una descripción interesante de las contribuciones de este creativo y competente individuo.
3. Lea una de las siguientes referencias sobre el método del meta-análisis.

Light, R. J. y Pillemer, D. B. (1984). *Summing up: The science of reviewing research.* Cambridge, Massachusetts: Harvard University Press.

Farley, J. U. y Lehmann, D. R. (1986). *Meta-analysis in marketing: Generalization of response models.* Lexington, Massachusetts: Lexington Books.

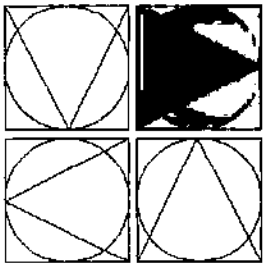
Plucker, J. A. (1997). Debunking the myth of the "highly significant" result: Effect sizes in gifted education research. *Roeper Review, 20*, 122-126.

Rosenthal, R. (1984). *Meta-analytic procedures for social research.* Thousand Oaks, California: Sage.

Sharpe, D. (1997). Of apples and oranges, file drawers and garbage: Why validity issues in meta-analysis will not go away. *Clinical Psychology Review, 17*, 881-901.

4. ¿Alguna vez ha sufrido de dolor de cabeza? Encontrará interesante el siguiente artículo.

McCrory, D. C. y Hasselblad, V. (1997). Cranial electrostimulation for headache: Meta-analysis. *Journal of Nervous and Mental Disease, 185*, 766-767.



CAPÍTULO 26

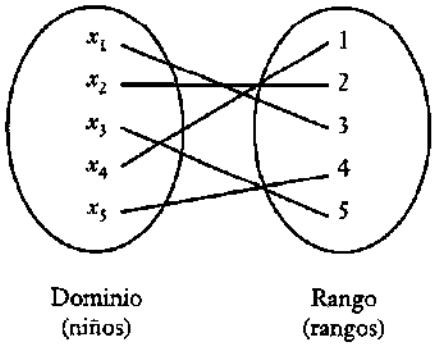
FUNDAMENTOS DE MEDICIÓN

- DEFINICIÓN DE MEDICIÓN
- ISOMORFISMO ENTRE MEDICIÓN Y "REALIDAD"
- PROPIEDADES, CONSTRUCTOS E INDICADORES DE OBJETOS
- NIVELES DE MEDICIÓN Y ESCALACIÓN
  - **Clasificación y enumeración**
  - **Medición nominal**
  - **Medición ordinal**
  - **Medición de intervalo (escalas)**
  - **Medición de razón (escalas)**
- COMPARACIÓN DE ESCALAS: CONSIDERACIONES PRÁCTICAS Y ESTADÍSTICOS

La medición es una de las piedras angulares de la investigación. Cualquier cuantificación de eventos, objetos, lugares y cosas involucra medición. Janda (1998) expresa acertadamente, en el prefacio de su libro, que la medición es fundamental para todas las áreas de la psicología y las ciencias sociales. Todos los procedimientos estadísticos descritos en este libro dependen de la medición. La mayoría de los métodos de recolección de datos, que eventualmente requieren algún tipo de cuantificación, se basan en la medición. Stevens (1951, 1968) afirma que "en su sentido más amplio, la medición es la asignación de valores numéricos a objetos o eventos, de acuerdo con ciertas reglas". La definición de Stevens expresa, de forma sucinta, la naturaleza básica de la medición. Para entenderla, sin embargo, se requiere definir y explicar cada término importante —tarea a la cual se dedica este capítulo—.

Suponga que se le pide a un juez que se pare a siete pies de distancia de un grupo de estudiantes, que observe a los estudiantes y que después estime el grado en que cada uno de ellos posee cinco atributos: simpatía, fuerza de carácter, personalidad, habilidad musical e inteligencia. Las estimaciones deben expresarse numéricamente con una escala de números del 1 al 5, donde el 1 indica una muy pequeña cantidad de la característica en cuestión, y 5 indica una gran cantidad de la misma. En otras palabras, con sólo observar a los estudiantes, el juez debe evaluar qué tan "simpáticos" son, qué tan "fuertes" son sus caracteres, etcétera, utilizando los números 1, 2, 3, 4 y 5 para indicar la cantidad de cada característica que ellos poseen.

FIGURA 26.2



la regla expresada en el párrafo anterior: "si una persona es mujer, asígnele un 1; si es hombre, asígnele un 0". Suponga que 0 y 1 forman el conjunto llamado B; entonces $B = \{0, 1\}$. El diagrama de medición se presenta en la figura 26.1.

Este procedimiento es igual al que se utilizó en el capítulo 5, cuando se discutió sobre relaciones y funciones. En efecto, la medición es una relación. Puesto que a cada miembro de A, el dominio, solamente se le asigna uno y solamente un objeto de B, el rango, la relación constituye una función. ¿Esto significa, entonces, que todos los procedimientos de medición son funciones? Sí, lo son, siempre que los objetos medidos sean considerados el dominio, y los valores numéricos a los que se asignen, o sobre los que se representen, sean considerados el rango.

Aquí hay otra forma de reunir los conceptos de conjunto, relación, función y medición. Recuerde que una relación es un conjunto de pares ordenados; también una función lo es. Entonces, cualquier procedimiento de medición establece un conjunto de pares ordenados, donde el primer miembro de cada par es el objeto medido, y el segundo miembro es el valor numérico asignado al objeto, en concordancia con la regla de medición, cualquiera que ésta sea. Así, ahora es posible escribir una ecuación general para cualquier procedimiento de medición:

$$f = \{(x, y);\ x = \text{cualquier objeto, y } y = \text{un valor numérico}\}$$

que se lee: "la función, f, o la regla de correspondencia, es igual al conjunto de pares ordenados (x, y), de tal manera que x es un objeto, y cada y correspondiente es un valor numérico". Se trata de una regla general que es adecuada para cualquier caso de medición.

Ahora se citará un ejemplo para hacer más concreto el análisis. Los eventos a medir, las x, son cinco niños. Los valores numéricos son los rangos 1, 2, 3, 4 y 5. Suponga que f es una regla que le indica a un maestro lo siguiente: "dé el rango 1 al niño que tenga la mayor motivación para hacer trabajo escolar. Dé el rango 2 al niño que tiene la siguiente mayor motivación para hacer trabajo escolar, y así sucesivamente, hasta el rango 5, el cual debe asignarse al niño que tenga la menor motivación para hacer trabajo escolar". La medición o la función aparece en la figura 26.2.

Observe que f, la regla de correspondencia, quizás habría sido: "si un niño tiene alta motivación para el trabajo escolar, déle un 1; pero si un niño tiene baja motivación para el trabajo escolar, déle un 0". Entonces, el rango sería {0, 1}. Ello tan sólo significa que el conjunto de cinco niños se ha dividido en dos subconjuntos, y a cada uno de ellos se les

asignará, por medio de *f*, los valores numéricos 0 y 1. Un diagrama de esto es similar a la figura 26.1, donde el conjunto *A* es el dominio y el conjunto *B* es el rango.

Regresando a las reglas, aquí es donde la evaluación entra en escena. Las reglas pueden ser "buenas" o "malas". Con reglas "buenas" se tiene una medición "buena" o acertada, si lo demás permanece igual. Con reglas "malas" se tiene una medición "mala" o pobre. Muchas cosas son fáciles de medir a causa de que las reglas son fáciles de elaborar y de seguir. Por ejemplo, medir el sexo resulta fácil, ya que varios criterios simples y bastante claros sirven para determinar el sexo y para indicar al investigador cuándo asignar 1 y cuándo asignar 0. También es fácil medir otras características humanas, tales como color de cabello, color de ojos, estatura o peso. Por desgracia, la mayoría de las características humanas son mucho más difíciles de medir, principalmente porque es difícil idear reglas claras que sean "buenas". No obstante, siempre deben tenerse reglas de algún tipo para medir cualquier cosa.

Isomorfismo entre medición y "realidad"

Como se ha visto, la medición puede ser un asunto sin sentido. ¿Cómo evitar esto? La definición de conjuntos de objetos a medir, la definición de conjuntos numéricos a partir de los cuales se asignan valores numéricos a los objetos que se miden, y las reglas de asignación o correspondencia, deben ligarse con la "realidad". Cuando se mide la dureza de objetos, hay poca dificultad. Si una sustancia *a* puede rayar a *b* (y no a la inversa), entonces *a* es más dura que *b*. De la misma forma, si *a* puede rayar a *b*, y *b* puede rayar a *c*, entonces (probablemente), *a* puede rayar a *c*. Éstas son cuestiones empíricas que son fáciles de comprobar, de tal manera que puede encontrarse un orden de rango de la dureza. Es posible medir la dureza de un conjunto de objetos por medio de unas cuantas pruebas de rayado, asignando valores numéricos para indicar el grado de dureza. Se afirma que el procedimiento de medición y el sistema de números son *isomórficos* a la realidad.

Isomorfismo significa identidad o similitud de forma. Las preguntas planteadas son: ¿este conjunto de objetos es isomórfico a aquel conjunto de objetos? ¿Los dos conjuntos son iguales o similares en algún aspecto formal? ¿Los procedimientos de medición utilizados tienen alguna correspondencia racional y empírica con la "realidad"?

Para demostrar la naturaleza del isomorfismo, es posible utilizar la idea de la correspondencia de conjuntos de objetos. Quizá se desea medir la *persistencia* de siete individuos. Suponga, también, que existe un ser omnisciente, que conoce la cantidad exacta de persistencia que cada individuo posee; es decir, conoce los valores "verdaderos" de persistencia de cada individuo. (Considere que *persistencia* ha sido definida adecuadamente.) Sin embargo, *usted*, quien mide, no conoce estos valores "verdaderos". Es necesario que usted *evalúe* la persistencia de los individuos de alguna forma falible, y usted piensa que ya encontró dicha forma. Por ejemplo, usted evaluaría la persistencia dándoles a los individuos una tarea que realizar y registrando el tiempo total que cada uno requiera para completarla, o puede anotar el número total de veces que el individuo intenta realizar la tarea antes de dirigirse a otra actividad (Feather, 1962). Usted utiliza su método y mide la persistencia de los individuos. Resultan, digamos, los siguientes siete valores: 6, 6, 4, 3, 3, 2, 1. Ahora, el ser omnisciente conoce los valores "verdaderos", que son 8, 5, 2, 4, 3, 3, 1. Este conjunto de valores es la "realidad". La correspondencia de su conjunto con la "realidad" se presenta en la figura 26.3.

En dos casos, usted ha evaluado los valores "verdaderos" de forma exacta; y ha "fallado" en todos los demás. Sin embargo, sólo una de estas "fallas" es seria, y hay una correspondencia bastante buena entre los dos órdenes de rango de los valores. Note, además,

dificultades surgen principalmente sobre el desacuerdo de los estadísticos que pueden utilizarse legítimamente para los diferentes niveles de medición. La posición de Stevens y la definición de medición citada anteriormente es una perspectiva amplia que, con relajación liberal, se sigue en este texto. Una posición más restrictiva —pero defendible— requiere que las diferencias entre las medidas puedan interpretarse como *diferencias cuantitativas de la propiedad medida*. En la perspectiva de algunos expertos, "cuantitativo" significa que una diferencia de magnitud entre dos valores de atributo representa una diferencia cuantitativa correspondiente en los atributos (véase Jones, 1971, pp. 335-355). Estrictamente hablando, esta visión excluye como *medición* a las escalas nominales y ordinales, las cuales se definirán en la siguiente sección de este capítulo. Los autores de este libro consideran que la experiencia real de medición en las ciencias del comportamiento y en la educación justifica una posición más relajada. Nuevamente, esto no tiene una importancia considerable, en caso de que el estudiante *entienda* las ideas generales presentadas. Se recomienda que el estudiante más avanzado lea los capítulos 1 y 2 de Torgerson (1958), y el capítulo 1 de Nunnally (1978); ambas referencias ofrecen buenas presentaciones. Comrey (1950, 1976) y Michell (1990) han influido de manera importante en la orientación que el segundo autor da a este capítulo. Comrey (1976) presenta un ensayo revelador sobre el problema fundamental de la medición en las ciencias sociales y del comportamiento. Un tratado más antiguo y excelente que ha ejercido gran influencia en esta obra es el de Guilford (1954). El estudiante curioso disfrutará la colección de artículos sobre la controversia publicada en el capítulo 2 de un libro editado por Kirk (1972). Los lectores que tengan intenciones de realizar investigación y que siempre se enfrentarán con problemas de medición deben leer cuidadosa y repetidamente las excelentes presentaciones que Nunnally (1978) o Nunnally y Bernstein (1994) hacen de los problemas y de su solución.

En el siguiente análisis, primero se considera el problema científico fundamental y de medición de la clasificación y la enumeración.

Clasificación y enumeración

El primer y más elemental paso en cualquier procedimiento de medición consiste en definir los objetos del universo de información. Suponga que U, el conjunto universal, se define como todos los alumnos de primer año de cierta preparatoria. A continuación, deben definirse las propiedades de los objetos de U. Todas las mediciones requieren que U se separe en, por lo menos, dos subconjuntos. La forma más elemental de medición sería clasificar o categorizar todos los objetos como poseedores o no de alguna característica. Considere que dicha característica es la condición masculina. Se separa U en hombres y no hombres, u hombres y mujeres. Éstos, por supuesto, son dos *subconjuntos* de U, o *particiones* de U. (Recuerde que partir un conjunto consiste en separarlo en subconjuntos que sean *mutuamente excluyentes* y *exhaustivos*; es decir, cada objeto del conjunto debe asignarse a uno y solamente un subconjunto, y que todos los objetos del conjunto de U deben asignarse de esta manera.)

Lo que se ha hecho es clasificar los objetos de interés. Se han ubicado en categorías: se han partido. La simpleza obvia de este procedimiento parece provocar dificultad a los estudiantes. La gente pasa gran parte de su vida categorizando cosas, eventos y personas. La vida no podría continuar sin dicha categorización, aunque asociar el proceso con la medición parece difícil de lograr.

Después de encontrar un método de clasificación, se tiene como efecto una regla que indica cuáles objetos de U van dentro de qué clases, subconjuntos o particiones. Se utiliza la regla y los objetos del conjunto se ubican en los subconjuntos. Aquí están los niños; acá

las niñas. Fácil. Aquí están los niños de clase media; acá los niños de clase trabajadora. No tan fácil, pero tampoco demasiado difícil. Aquí están los delincuentes; acá los no delincuentes. Más difícil. Aquí están los destacados; acá los mediocres, y más allá los lerdos. Mucho más difícil. Aquí están quienes son creativos; acá quienes no son creativos. Muchísimo más difícil.

Después de que los objetos del universo se han clasificado dentro de subconjuntos designados, es posible contar a los miembros de los conjuntos. En caso de dicotomía, la regla de conteo fue expresada en el capítulo 4. Si un miembro de U posee la característica en cuestión, por ejemplo, condición masculina, entonces se asigna 1. Si el miembro no posee la característica, entonces se asigna 0 (véase figura 26.1). Cuando los miembros del conjunto se cuentan de esta manera, todos los objetos de un subconjunto se consideran iguales entre sí, y desiguales respecto a los miembros de otros subconjuntos.

Existen cuatro niveles generales de medición: nominal, ordinal, de intervalo y de razón. Estos cuatro niveles conducen a cuatro tipos de escalas. Algunos escritores sobre el tema aceptan únicamente la medición ordinal, de intervalo y de razón; mientras que otros afirman que los cuatro pertenecen a la familia de la medición. Comrey y Lee (1995) consideran que la escala nominal constituye una forma de medición. Sin embargo, ésta no es tan cuantitativa como la ordinal, la de intervalo y la de razón. Es decir, los números utilizados en la medición nominal son sólo etiquetas numéricas ligadas a categorías predefinidas. No es necesario ser tan exigentes respecto a esto mientras se comprendan las características de las diferentes escalas y niveles.

Medición nominal

Las reglas utilizadas para asignar valores numéricos a los objetos definen el tipo de escala y el nivel de medición. El nivel más bajo de medición es el *nominal* (véase el análisis previo sobre categorización). Los números asignados a los objetos son valores numéricos que no tienen un significado numérico; no pueden ordenarse o sumarse. Son *etiquetas*, parecidas a las letras que se utilizan para nombrar conjuntos. Si a grupos o individuos se les asigna 1, 2, 3, tales valores numéricos son simplemente nombres. Por ejemplo, a los jugadores de beisbol y de futbol se les asignan este tipo de números; a los teléfonos también. A los grupos se les pueden asignar las etiquetas I, II y III o A_1, A_2 y A_3. Utilizamos medición nominal en nuestro pensamiento y vida cotidianos. Identificamos a otros como "hombres", "mujeres", "protestantes", "australianos", etcétera. De cualquier manera, los símbolos asignados a objetos, o mejor dicho, a conjuntos de objetos, constituyen escalas nominales. Algunos expertos no creen que esto sea medición, como se indicó previamente. Pero dicha exclusión de la medición nominal no permitiría que muchos de los procedimientos de investigación en ciencias sociales fuesen llamados medición. Puesto que se satisface la definición de medición y como los miembros de los conjuntos etiquetados pueden contarse y compararse, parece que los procedimientos nominales *son* medición.

Los requisitos de la medición nominal son simples. A todos los miembros de un conjunto se les asigna el mismo valor numérico, y no se le asigna el mismo valor numérico a dos conjuntos. La medición nominal —al menos en una forma simple— fue expresada en la figura 26.1, donde los objetos del rango {0, 1} quedaron representados en las a, los objetos de U, las cinco personas, por medio de la regla: "si x es hombre, asignar 0; si x es mujer, asignar 1". Ésta es la manera en que se cuantifica la medición nominal cuando está involucrada únicamente una dicotomía. Cuando la partición contiene más de dos categorías, debe utilizarse algún otro método. La cuantificación de medición nominal básicamente equivale a contar objetos en las casillas de los subconjuntos o particiones.

considera bueno desde el punto de vista de la medición (escalación). La conversión de estas mediciones a puntuaciones estándar o Z, resulta en unidades que pueden considerarse cuantitativamente iguales. Los métodos de escalación que utilizan la curva normal para obtener mediciones en la escala de intervalo pueden, cuando mucho, considerarse aproximaciones con precisión desconocida. Comrey y Lee (1995) presentan un método de este tipo en el capítulo 5 de su libro.

Medición de razón (escalas)

El nivel más alto de medición es el de razón, y el ideal de medición de los científicos es la escala de razón. Una *escala de razón*, además de poseer las características de las escalas nominal, ordinal y de intervalo, posee un cero absoluto o natural con significado empírico. Si una medición es cero en una escala de razón, entonces existe una base para afirmar que un objeto no posee la característica medida. Puesto que existe un cero absoluto o natural, es posible realizar todas las operaciones aritméticas, incluyendo la multiplicación y la división. Los números de la escala indican las cantidades reales de la propiedad medida. Si existiera una escala de razón del *rendimiento*, entonces sería posible decir que un alumno con una puntuación de 8 en la escala posee un rendimiento dos veces mayor que un alumno con una puntuación de 4 en la misma escala.

Uno de los principales problemas en las ciencias sociales y del comportamiento es que la operación de suma no puede definirse (Comrey, 1950). Además, no existen sustitutos satisfactorios reales para el operador de suma en las ciencias sociales y del comportamiento que permita al investigador obtener una escala de medición de razón. Hubo algunos procedimientos de escalación que fueron complejos y parcialmente exitosos, pero, en general, los datos con los que trabajan los científicos sociales y del comportamiento no son siquiera aproximadamente cercanos a datos de una escala de razón.

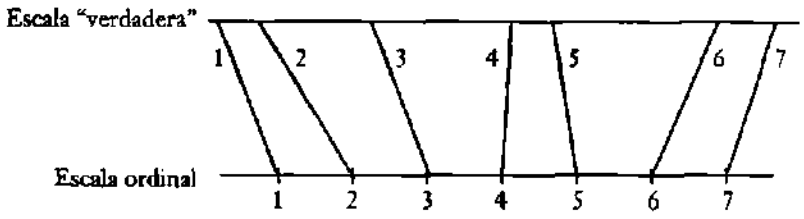
Comparación de escalas: consideraciones prácticas y estadísticos

Las características básicas de los cuatro tipos de medición y sus escalas acompañantes ya se han analizado. ¿Qué tipo de escalas se utilizan en la investigación educativa y del comportamiento? Se utilizan principalmente la nominal y la ordinal, aunque existe una alta posibilidad de que muchas escalas y pruebas utilizadas en la medición psicológica y educativa se aproximen a la medición de intervalo lo suficientemente para propósitos prácticos, como se verá más adelante.

Primero, considere la medición nominal. Cuando los objetos se dividen en dos, tres o más categorías con base en la pertenencia a un grupo —sexo, identificación étnica, casado-soltero, protestante-católico-judío, etcétera— la medición es nominal. Cuando las variables continuas se convierten en atributos, como cuando los objetos se dividen en alto-bajo y viejo-joven, se obtiene lo que puede llamarse medición cuasi-nominal: aunque sujetos de, por lo menos, un orden de rango, los valores son, en efecto, colapsados a 1 y 0.

Resulta instructivo estudiar las operaciones numéricas que son, en un sentido estricto, legítimas con cada tipo de medición. En la medición nominal se permite, por supuesto, el conteo del número de casos en cada categoría y subcategoría. Los estadísticos de frecuencia, como los porcentajes de χ^2 y ciertos coeficientes de correlación (coeficientes de contingencia) pueden utilizarse. Esto suena poco; pero en realidad es bastante. Un buen principio que debe recordarse es éste: si no es posible utilizar cualquier otro método, casi

FIGURA 26.4



siempre es posible realizar una partición cruzada con los participantes. Si se estudia la relación entre dos variables y no se tiene forma adecuada de medirlos de manera ordinal o de intervalo, quizá se pueda encontrar una forma de dividir los objetos de estudio en por lo menos dos grupos. Por ejemplo, al estudiar la relación entre la motivación de los miembros de un consejo de educación para convertirse en miembros del consejo y su religión, como lo hicieron Gross, Mason y McEachern (1958), se pide a jueces expertos que dividan la muestra de miembros del consejo en aquellos con "buena" motivación y aquellos con motivación "pobre". Después se puede hacer una partición cruzada de la religión respecto a la dicotomía de motivación, y así estudiar la relación.

Las puntuaciones de pruebas de inteligencia, aptitud y personalidad son, *hablando de forma básica y estricta*, ordinales. Éstas indican de forma más o menos precisa, no las *cantidades* de los rasgos de inteligencia, aptitud y personalidad de los individuos, sino más bien las *posiciones del orden de rango* de los individuos. Para verlo, es necesario darse cuenta de que las escalas ordinales no poseen las características deseables de igualdad de intervalos o ceros absolutos. Las puntuaciones de pruebas de inteligencia constituyen algunos ejemplos. Un individuo con una puntuación de cero en una medida de inteligencia no necesariamente carece de ella, ya que no existe un cero absoluto en la escala de una prueba de inteligencia. El cero es arbitrario y al no tener un cero absoluto la suma de *cantidades* de inteligencia no tiene ningún sentido, puesto que los puntos de cero arbitrarios conducen a sumas diferentes. Sumar a dos personas cuando cada una tiene una puntuación de inteligencia de 70 no es equivalente a una persona con un CI de 140. En una escala con un punto cero arbitrario se realiza la siguiente suma: 2 + 3 = 5. Entonces, la suma es 5 unidades escalares por arriba de cero. Pero si el punto cero arbitrario es impreciso y el punto del cero "real" está 4 puntos más abajo que la posición del cero arbitrario de la escala, entonces los anteriores 2 y 3 en realidad deberían ser 6 y 7, ¡y 6 + 7 = 13!

La falta de un cero real en las escalas ordinales no es tan seria como la falta de intervalos iguales. Aun sin un cero real, pueden añadirse *distancias* dentro de la escala, siempre y cuando tales distancias sean iguales (empíricamente). La situación podría parecerse a la indicada en la figura 26.4. La escala en la parte superior (escala "verdadera") indica los valores "verdaderos" de una variable. La escala de la parte inferior (escala ordinal) indica la escala de orden de rango utilizada por un investigador. En otras palabras, un investigador ha ordenado por rango a siete personas bastante bien; pero sus valores numéricos ordinales, que se ven con intervalos iguales, no son "verdaderos", aunque puedan ser representaciones bastante precisas de los hechos empíricos.

Estrictamente hablando, los estadísticos que pueden utilizarse con escalas ordinales incluyen las medidas de orden de rango, tales como el coeficiente de correlación de orden de rango, r, la W de Kendall y el análisis de varianza de orden de rango, las medianas y los percentiles. Si únicamente son legítimos dichos estadísticos (y otros similares), ¿cómo es

En el estado que guarda actualmente la medición, no se puede estar seguro de que los instrumentos de medición tengan intervalos iguales. Es importante plantear la pregunta: ¿qué tan serias son las distorsiones y errores introducidos al tratar las mediciones ordinales como si fueran mediciones de intervalo? Al tener cuidado en la construcción de instrumentos de medición, y especial cuidado en la interpretación de los resultados, las consecuencias evidentemente no son serias. Los métodos estadísticos más poderosos dependen menos de la escala de medición subyacente que de las propiedades de distribución de los datos.

El mejor procedimiento parecería ser tratar las mediciones ordinales como si fueran mediciones de intervalo; pero estando constantemente alertas a la posibilidad de desigualdades grandes en los intervalos. Debe aprenderse lo más posible acerca de las características de las herramientas de medición. A través de la apropiada refinación de los métodos de medición y de los procedimientos de escalación, es posible obtener datos que sean aproximadamente normales en su forma. Con datos de este tipo, se pueden utilizar métodos paramétricos de análisis estadístico más poderosos. El investigador debe estar consciente de que es incorrecto ignorar las propiedades escalares de los datos. Por ejemplo, sería inapropiado que un investigador interpretara un grupo con una media de 50 como el doble de un grupo que tuviera una media de 25. Mucha información útil se ha obtenido al tratar datos ordinales como de intervalo, lo que ha resultado en avances científicos en psicología, sociología y educación. En pocas palabras, es muy improbable que los investigadores sean conducidos por mal camino al seguir este consejo, si son cuidadosos al aplicarlo. Para encontrar una útil revisión de la literatura sobre el problema de las escalas de medición y estadística, revise Gardner (1975) o Michell (1990).

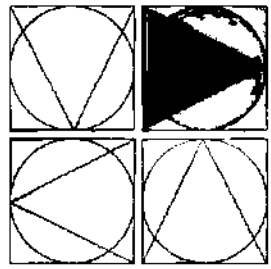
RESUMEN DE CAPÍTULO

1. La medición es un componente importante de investigación.
2. Sin medición o cuantificación de información, muchos métodos de análisis estadístico no podrían utilizarse.
3. Stevens define la medición como el proceso de asignación de números a objetos y eventos, de acuerdo con alguna regla.
4. Stevens define cuatro conjuntos de reglas: nominal, ordinal, de intervalo y de razón.
5. La mayor parte de los datos de las ciencias sociales y del comportamiento son ordinales. Sin embargo, a través de ciertos métodos y supuestos de escalación, pueden considerarse como datos de escala de intervalo.
6. Comrey afirma que una consideración importante es que los datos de las ciencias del comportamiento pueden considerarse de intervalo, si el proceso de medición genera datos que tengan una distribución normal.
7. La medición implica un isomorfismo entre los números y la realidad.
8. Continúa la discusión sobre cuál es la mejor forma de manejar datos de las ciencias sociales y del comportamiento.

SUGERENCIAS DE ESTUDIO

1. ¿Cuál es el primer paso en la medición?
2. De acuerdo a Stevens, ¿cuáles son las reglas que forman parte del proceso de medición?

3. Dé un ejemplo de la ciencia o de la vida diaria que ilustre la medición ordinal.
4. Un artículo interesante escrito hace muchos años por Prokasy (1962) es relevante aun para la discusión actual sobre el uso de métodos paramétricos para datos ordinales. Lea el artículo de Prokasy y, después, revise el capítulo 1 de Cliff (1996).
5. Lea el artículo de F. M. Lord sobre el tratamiento estadístico de datos de futbol americano (en Kirk, 1972). En él se describe, de manera humorística, cómo la gente percibe y utiliza los números. ¿Los números de una escala nominal pueden sumarse?



CAPÍTULO 27

CONFIABILIDAD

- DEFINICIONES DE CONFIABILIDAD
- TEORÍA DE LA CONFIABILIDAD
  - Dos ejemplos computacionales
- INTERPRETACIÓN DEL COEFICIENTE DE CONFIABILIDAD
- EL ERROR ESTÁNDAR DE LA MEDIA Y EL ERROR ESTÁNDAR DE MEDICIÓN
- INCREMENTO DE LA CONFIABILIDAD
- EL VALOR DE LA CONFIABILIDAD

Después de asignar valores numéricos a los objetos o eventos de acuerdo con reglas, deben enfrentarse dos grandes problemas de medición: la confiabilidad y la validez. Ya se ha diseñado un sistema de medición y se han administrado los instrumentos de medición a un grupo de participantes. Ahora deben preguntarse y responderse las siguientes preguntas: ¿cuál es la confiabilidad del instrumento de medición? ¿Cuál es su validez?

Si no se conoce la confiabilidad ni la validez de los propios datos, es posible que haya poca fe en los resultados obtenidos y en las conclusiones obtenidas a partir de ellos. Éstas son dos propiedades psicométricas clave que deben ser satisfechas para responder a las muchas críticas hechas a los datos de las ciencias sociales y del comportamiento, así como a los métodos de medición. Los datos de las ciencias sociales y de educación, derivados de la conducta humana y de productos humanos están, como se vio en el capítulo 26, un poco alejados de las propiedades del interés científico; por lo tanto, su validez puede cuestionarse. La preocupación por la confiabilidad proviene de la necesidad de fiarse de la medición. Los datos provenientes de todos los instrumentos de medición en psicología y educación contienen errores de medición. Dependiendo del grado en que contengan errores, los datos que produzcan serán fiables o no.

Definiciones de confiabilidad

Sinónimos de confiabilidad son *estabilidad*, *fiabilidad*, *consistencia*, *reproductibilidad*, *predictibilidad* y *falta de distorsión*. Por ejemplo, las personas confiables son aquellas cuyo

comportamiento es consistente, predecible y fiable; lo que hacen mañana y la siguiente semana será consistente con lo que hacen hoy y con lo que hicieron la semana pasada; se dice que son estables. Por otro lado, las personas poco confiables son aquellas cuyo comportamiento es mucho más variable; son impredeciblemente variables. En algunas ocasiones hacen algo; y en otras, algo distinto; carecen de estabilidad. Se dice que son inconsistentes.

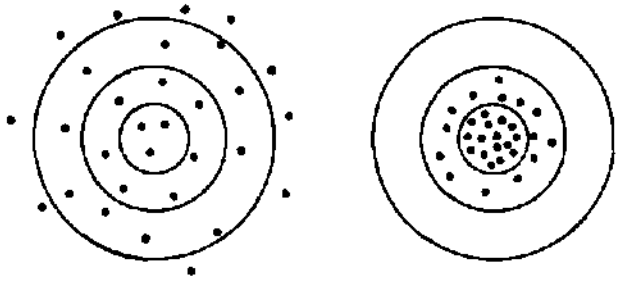
Lo mismo sucede con las mediciones en psicología y en educación: son más o menos variables de una ocasión a otra. O son estables o relativamente predecibles, o son inestables y relativamente impredecibles; son consistentes o no lo son. Si son confiables, entonces se puede depender de ellas; si no son confiables, no se puede depender de ellas.

La definición de confiabilidad se enfoca de tres maneras: un enfoque se sintetiza con la pregunta: si se mide el mismo conjunto de objetos una y otra vez, con el mismo instrumento de medición o uno comparable, ¿se obtendrán iguales o similares resultados? La pregunta implica una definición de confiabilidad en términos de *estabilidad, fiabilidad* y *predictibilidad.* Es la definición que se ofrece en discusiones elementales del tema.

Un segundo enfoque se sintetiza con la pregunta: ¿las medidas obtenidas a partir de un instrumento de medición son las medidas "verdaderas" de la propiedad que se mide? Ésta es una definición de *falta de distorsión*. Comparada con la primera definición, se aleja más del sentido común y de la intuición; sin embargo, es también más fundamental. Estos dos enfoques o definiciones se resumen en las palabras *estabilidad* y *falta de distorsión*. Sin embargo, como se verá más adelante la definición sobre la falta de distorsión implica la definición de estabilidad. La confiabilidad se refiere al grado en el que la medición concuerda consigo misma. En el capítulo 28 se tratará la validez. Con frecuencia los términos "confiabilidad" y "validez" se confunden, no obstante existe una clara distinción entre ellos. La confiabilidad no tiene nada que ver con la veracidad de la medición. Algunos autores se han referido a la confiabilidad como precisión (véase Magnusson, 1967; Tuckman, 1975). Esto es verdad, pero con frecuencia se confunde con el significado de precisión en términos de validez. La validez también tiene que ver con la precisión, pero de una manera diferente que la confiabilidad. La confiabilidad se relaciona con la precisión con la que un instrumento de medición mide aquello que se desea. La palabra clave aquí es "aquello". Si se tiene una prueba que se considera que mide habilidad matemática, no se sabe si la prueba mide, en realidad, habilidad matemática. Si la prueba es altamente confiable, solamente se sabe que está midiendo "algo" con precisión. El asegurarse de que la prueba de habilidad matemática en realidad mide habilidad matemática, implica involucrarse con aspectos de validez.

Existe un tercer enfoque en la definición de confiabilidad, el cual no sólo ayuda a lograr una mejor definición y a resolver tanto problemas teóricos como prácticos, sino que también implica otros enfoques y definiciones. Se puede investigar qué tanto *error de medición* existe en un instrumento de medición. Recuerde que existen dos tipos generales de varianza: sistemática y por el azar. La *varianza sistemática* se inclina hacia una dirección —las puntuaciones tienden a ser todas negativas o todas positivas, o todas altas o todas bajas—. En este caso el error es constante o está sesgado. La *varianza por el azar* o *del error* se autocompensa— las puntuaciones tienden a inclinarse ahora hacia este lado, ahora hacia este otro—. Los errores de medición son errores aleatorios; representan la suma de diversas causas. Entre dichas causas están los elementos comunes del azar o aleatorios —presentes en todas las medidas debido a causas desconocidas—, la fatiga temporal o momentánea, las condiciones fortuitas que en un momento en particular afectan al objeto medido o al instrumento de medición, las fluctuaciones en la memoria y en el estado de ánimo, y otros factores que son temporales y cambiantes. Dependiendo del grado en que los errores de medición estén presentes en un instrumento de medición, el instrumento

FIGURA 27.1

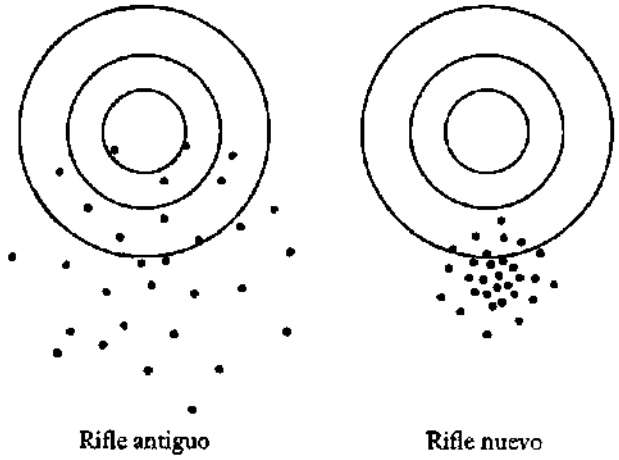


será poco confiable. En otras palabras, la confiabilidad puede definirse como la ausencia relativa de errores de medición en un instrumento de medición.

La confiabilidad es la *falta de distorsión* o *precisión* de un instrumento de medición. Recuerde que una medida altamente confiable sólo indica que está midiendo algo con precisión o de forma consistente. Puede ocurrir que no esté midiendo lo que se cree que mide. Un ejemplo para ilustrar lo anterior es la báscula que tenemos en nuestros hogares. Suponga que esta báscula siempre sobrestima el peso de una persona por 5 kilogramos. Si alguien se coloca sobre esta báscula 50 veces durante el periodo de una hora, encontrará muy poca fluctuación del peso registrado en la báscula. La báscula es precisa en el sentido de que indica consistentemente el mismo peso. Sin embargo, es imprecisa en el sentido de que siempre da un peso equivocado por 5 kilogramos. La báscula sería considerada confiable, pero no válida.

Considere que un deportista desea comparar la precisión de dos armas. Una es una pieza antigua fabricada hace un siglo, pero que se encuentra aún en buenas condiciones. La otra es un arma moderna fabricada por un armero experto. Ambas piezas se encuentran fijas en bases de granito y son accionadas hacia un blanco por un pistolero experto. Cada

FIGURA 27.2



arma se dispara igual número de veces. En la figura 27.1 se presenta el patrón hipotético de tiros a un blanco para cada una. El blanco de la izquierda representa el patrón de tiros producido por el arma antigua; observe que los tiros se encuentran considerablemente dispersos. Ahora considere que el patrón de tiros en el blanco de la derecha está más junto. Los tiros se encuentran agrupados de forma cercana alrededor del blanco.

Suponga que se asignan números a los círculos del blanco: 3 al centro, 2 al círculo siguiente, 1 al círculo externo y 0 a cualquier tiro que salga del blanco. Es obvio que si se calculan medidas de variabilidad, por ejemplo, una desviación estándar, de los dos patrones de tiro, el rifle antiguo tendría una medida de variabilidad mucho más grande que el rifle más nuevo. Estas medidas pueden considerarse índices de confiabilidad. La medida menor de variabilidad del rifle nuevo indica mucho menos error y, por lo tanto, mucho mayor precisión. El rifle nuevo es confiable; el rifle antiguo es menos confiable.

Ahora analice la figura 27.2. Aquí se tiene el mismo patrón de tiros de ambos rifles; aunque no están centrados en el blanco como en la figura 27.1. El rifle nuevo seguiría considerándose más confiable que el antiguo, pero debido a que ambos se salen del blanco, entonces la puntería no es precisa. Aquí los patrones de la precisión de los tiros de los rifles miden confiabilidad; mientras que la precisión de la puntería de los rifles mide validez. La figura 27.1 ilustra una manera burda de demostrar confiabilidad con validez; en cambio, la figura 27.2 demuestra confiabilidad con poca o ninguna validez. Es posible tener confiabilidad sin validez, pero no a la inversa. La confiabilidad por sí misma resulta poco útil para evaluar la mayoría de las mediciones. Como se indicó antes, una medición puede ser errónea consistentemente. No existe garantía de que el instrumento de medición sea bueno. No obstante, la ausencia de una confiabilidad alta sí indica que el instrumento de medición es pobre.

De forma similar, las mediciones en psicología y educación poseen mayores y menores confiabilidades. Se aplica un instrumento de medición, por ejemplo, una prueba de rendimiento aritmético, a un grupo de niños —generalmente sólo una vez—. La meta, por supuesto, es múltiple: se busca obtener la puntuación "verdadera" de cada niño. En la medida en que se fallen las puntuaciones "verdaderas", el instrumento de medición, la prueba, resulta poco confiable. Las puntuaciones aritméticas "verdaderas" y "reales" de cinco niños, por ejemplo, son 35, 31, 29, 22, 14. Otro investigador desconoce estas puntuaciones "verdaderas". Los resultados obtenidos son 37, 30, 26, 24, 15. Aunque en ningún caso se logró la puntuación "verdadera", todas poseen el mismo orden de rango. La confiabilidad y precisión del investigador son sorprendentemente altas.

Suponga que las cinco puntuaciones hubiesen sido 24, 37, 26, 15, 30. Éstas son las mismas cinco puntuaciones; aunque presentan un orden de rango muy diferente. En este caso, la prueba no sería confiable a causa de su falta de precisión. Para demostrar esto de

TABLA 27.1 *Puntuaciones y órdenes de rango "verdaderos", confiables y no confiables obtenidos de cinco niños*

(1) Puntuaciones "verdaderas"	(Rango)	(2) Puntuaciones de una prueba confiable	(Rango)	(3) Puntuaciones de una prueba no confiable	(Rango)
35	(1)	37	(1)	24	(4)
31	(2)	30	(2)	37	(1)
29	(3)	26	(3)	26	(3)
22	(4)	24	(4)	15	(5)
14	(5)	15	(5)	30	(2)

forma más compacta, los tres conjuntos de puntuaciones, con sus órdenes de rango, se han colocado unos junto a otros en la tabla 27.1. Las órdenes de rango de la primera y segunda columnas covarían de manera exacta. El coeficiente de correlación del orden de rango es 1.00. Aun cuando las puntuaciones de la prueba de la segunda columna no son exactas, se encuentran en el mismo orden de rango. Con base en esto, por medio del uso de un coeficiente de correlación del orden de rango, la prueba es confiable. Sin embargo, el coeficiente de correlación entre los rangos de la primera y tercera columnas es cero, de tal modo que la última prueba no es confiable por completo.

Teoría de la confiabilidad

El ejemplo presentado en la tabla 27.1 sintetiza lo que se debe saber acerca de la confiabilidad. El tratamiento que en este capítulo se da a la confiabilidad está basado en la teoría clásica de las pruebas. Existe un tratamiento mucho más avanzado de confiabilidad realizado por Cronbach, Gleser, Nanda y Rajaratnam (1972), llamado teoría de generalización. Aquí se tratará el modelo más tradicional de confiabilidad. Para hacerlo, es necesario formalizar los conceptos intuitivos y describir una teoría de la confiabilidad, la cual no sólo es elegante conceptualmente, sino que también es poderosa prácticamente. Resulta útil unificar las ideas sobre medición y proporciona un fundamento para comprender varias técnicas analíticas. La teoría también se relaciona de forma adecuada con el modelo de varianza enfatizado en análisis previos.

Cualquier conjunto de medidas posee una varianza total; es decir, después de aplicar un instrumento a un conjunto de objetos y de obtener un conjunto de números (puntuaciones), es posible calcular una media, una desviación estándar y una varianza. Aquí solamente se tratará la varianza, la cual, como se vio antes, es una varianza total obtenida, ya que incluye varianzas debidas a múltiples causas. En general, cualquier *varianza total obtenida* (o suma de cuadrados) incluye la varianza sistemática y del error.

Cada persona posee una puntuación obtenida, X_t. (La "*t*" significa "total".) Algunos autores se refieren a ella como la puntuación observada. Algunas ocasiones sólo se anota "*O*" o X_o. Ésta sería la medición que se hace de un objeto, persona, cosa o evento. La puntuación observada tiene dos componentes: un componente "verdadero" y un componente de error. Se supone que cada persona tiene una puntuación "verdadera", X_∞. (El símbolo "∞" de infinito se utiliza para representar lo "verdadero".) Un símbolo alternativo que el lector puede encontrar en la literatura es T o X_T. Dicha puntuación sería conocida sólo por un ser omnisciente, porque el sistema de medición es imperfecto. Note además lo que se estableció anteriormente. La puntuación verdadera puede incluir propiedades diferentes de la propiedad que se desea medir. El problema para medir esa propiedad es de validez. El otro componente es la puntuación de error, X_e o E; en este caso, error no significa un error que se haya cometido, sino que la puntuación de error es algún incremento o decremento que resulta de varios de los factores responsables de la incapacidad para medir la puntuación verdadera. Por ejemplo, un estudiante quizá tenga una puntuación observada menor que la puntuación "verdadera" debido a que esa persona estuvo enferma el día del examen. Por lo tanto, se puede afirmar que la diferencia entre la puntuación real y la observada es un error. Algunos errores son contables y otros no lo son.

La lógica conduce a una ecuación básica simple para la teoría:

$$X_t = X_\infty + X_e \qquad \textbf{(27.1)}$$

o

$$X_o = X_T + X_e$$

o

$$O = T + E$$

Esto establece, de forma sucinta, que cualquier puntuación observada está formada de dos componentes: un componente "verdadero" y un componente de error. La única parte de esta definición que representa un problema real es X_∞, que se concibe como la puntuación que un individuo obtendría si todas las condiciones internas y externas fueran "perfectas" y si el instrumento de medición fuera también "perfecto". De manera más realista se considera que es la media de un gran número de aplicaciones de la prueba a la misma persona. Simbólicamente, $X_\infty = (X_1 + X_2 + \ldots + X_n)/n$. Lord y Novick (1968) llaman a la puntuación "verdadera" el valor esperado de una puntuación observada, el cual puede interpretarse como la puntuación promedio que un individuo obtendría si toma un número infinito de mediciones independientes repetidas. Considérese lo siguiente: si una persona deseara conocer su estatura, ella puede medirse una vez. ¿Dará esto su estatura "verdadera"? Es poco probable, ya que el aparato de medición es falible. Por lo tanto, la persona haría bien en tomar múltiples mediciones de su estatura y, después, calcular la media de las estaturas. Esta media estaría más cerca de su estatura verdadera que cualquier medición hecha de forma aislada. Si el número de mediciones se acerca al infinito, la media se iría acercando cada vez más a la estatura verdadera.

Con un poco de álgebra simple, la ecuación 27.1 se extiende para producir una ecuación más útil en términos de varianza:

$$V_T = V_\infty + V_E \tag{27.2}$$

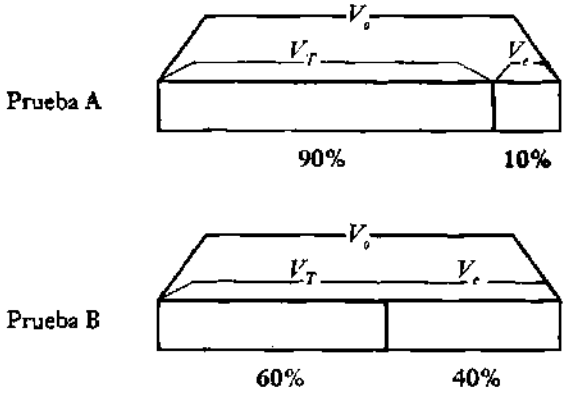
o

$$V_O = V_T + V_e$$

La ecuación 27.2 indica que la varianza total obtenida, de una prueba, se forma de dos componentes de varianza: un componente "verdadero" y un componente de "error". Si, por ejemplo, fuese posible aplicar el mismo instrumento al mismo grupo 4 367 929 veces, y después calcular las medias de las 4 367 929 puntuaciones de cada persona, se tendría un conjunto de mediciones "casi verdaderas" del grupo. En otras palabras, estas medias son las X_∞ del grupo. Entonces podría calcularse la varianza de las X_∞, produciendo V_∞. Este valor siempre debe ser menor que V_t o V_0, la varianza calculada a partir del conjunto de puntuaciones originales obtenido (las X_t u O), debido a que las puntuaciones originales contienen error. Sin embargo, las puntuaciones "verdaderas" o "casi verdaderas" no poseen error, ya que éste se ha eliminado por medio del proceso del cálculo de promedios. En otras palabras, si no hubiese errores de medición en las X_t u O, entonces $V_t = V_\infty$ o $V_0 = V_T$. Pero siempre existen errores de medición, y se supone que si se conocieran las puntuaciones de error y se restaran de las puntuaciones obtenidas, entonces se obtendrían las puntuaciones "verdaderas".

Nunca se conocen las puntuaciones "verdaderas" ni tampoco se conocen realmente las puntuaciones de error. No obstante, es posible estimar la varianza del error y, al hacerlo, en efecto es posible sustituir la ecuación 27.2 y resolverla. Ésta es la esencia de la idea, aunque se han omitido ciertos supuestos y pasos de la discusión. Un diagrama muestra las ideas de forma más clara. Sean las varianzas totales de dos pruebas representadas por medio de dos barras. Una prueba es altamente confiable; la otra lo es sólo moderadamente, como se indica en la figura 27.3. Las pruebas A y B tienen la misma varianza total, pero el 90% de la prueba A es varianza "verdadera" y el 10% es varianza del error. Únicamente el 60% de la prueba B es varianza "verdadera" y el 40% restante varianza del error. Por lo tanto, la prueba A es mucho más confiable que la prueba B.

FIGURA 27.3



La confiabilidad se define, por decirlo de alguna manera, a través del error; a mayor error, menor confiabilidad; y a menor error, mayor confiabilidad. Hablando de forma práctica, lo anterior significa que si se estima la varianza del error de una medida, entonces también se puede estimar la confiabilidad de la medida, lo cual conduce a dos definiciones de confiabilidad equivalentes:

1. La confiabilidad es la proporción de la varianza "verdadera" respecto de la varianza total obtenida de los datos producidos por un instrumento de medición.
2. La confiabilidad es la proporción de la varianza del error respecto de la varianza total producida por un instrumento de medición, restado de 1.00; donde el índice 1.00 indica una confiabilidad perfecta.

Resulta más fácil escribir las definiciones en forma de ecuación:

$$r_{tt} = \frac{V_{\infty}}{V_t} = \frac{V_T}{V_O} \tag{27.3}$$

$$r_{tt} = 1 - \frac{V_e}{V_t} = 1 - \frac{V_e}{V_o} \tag{27.4}$$

donde r_{tt} es el coeficiente de confiabilidad y los otros símbolos fueron ya definidos antes. La ecuación 27.3 es teórica y no puede utilizarse para realizar cálculos. La ecuación 27.4 es tanto teórica como práctica; se utiliza tanto para conceptualizar la idea de confiabilidad como para estimar la confiabilidad de un instrumento. Una ecuación alternativa a (27.4) es:

$$r_{tt} = \frac{V_t - V_e}{V_t} = \frac{V_o - V_e}{V_o} \tag{27.5}$$

Esta ecuación alternativa de la confiabilidad será útil para ayudar a comprender lo que es la confiabilidad.

Dos ejemplos computacionales

Para mostrar la naturaleza de la confiabilidad, en la tabla 27.2 se muestran dos ejemplos. Uno, denominado I en la tabla, es un ejemplo de alta confiabilidad; el otro, denominado II, es un ejemplo de baja confiabilidad. Note con cuidado que se utilizan exactamente los mismos números en ambos casos. La única diferencia es que están ordenados de manera distinta. La situación en ambos casos es: a cinco individuos se les aplicó una prueba con cuatro reactivos. (Lo cual es poco realista, por supuesto, aunque ayudará a ilustrar varias cuestiones.) Los datos de los cinco individuos se encuentran en los renglones; las sumas de los individuos se muestran a la derecha de los renglones (Σ_i). Las sumas de los reactivos se presentan en la parte inferior de cada tabla (Σ_{it}). Además, las sumas de los individuos en los reactivos impares (Σ_{impar}) y las sumas de los individuos de los reactivos pares (Σ_{par}) se presentan en la extrema derecha de cada subtabla. Los cálculos necesarios para el análisis de varianza de dos factores se muestran debajo de las tablas de datos.

Para volver estos ejemplos más realistas, imagine que los datos son puntuaciones en una escala de 6 puntos respecto a, por ejemplo, las actitudes hacia la escuela. Una puntuación elevada significa una actitud altamente favorable; una puntuación baja, una actitud poco favorable (o nada favorable). (Sin embargo, no hace ninguna diferencia cuáles son las puntuaciones. Inclusive pueden ser unos y ceros resultantes de marcar los reactivos de una prueba de rendimiento: correcto es igual a 1, e incorrecto es igual a 0.) En I, el individuo 1 tiene una actitud altamente favorable hacia la escuela; mientras que el individuo 5 tiene una actitud poco favorable hacia la escuela. Éstas ya están indicadas por las sumas de los individuos (o las medias): 21 y 5. Dichas sumas (Σ_i) son las puntuaciones comúnmente producidas por pruebas. Por ejemplo, si se quisiera conocer la media del grupo, se calcularía como $(21 + 18 + 14 + 10 + 5)/5 = 13.60$.

La varianza de estas sumas proporciona uno de los términos de las ecuaciones 27.4 y 27.5, pero no el otro: V_t pero no V_r. Utilizando el análisis de varianza es posible calcular tanto V_t como V_r. Los análisis de varianza de I y II indican cómo se hace esto. No es necesario ocuparse demasiado de estos cálculos, ya que son secundarios al tema principal.

El análisis de varianza produce las varianzas: entre reactivos, entre individuos y residual o del error. Las razones F de los reactivos no son significativas en I ni en II. (Observe que ambos cuadrados medios son 2.27. Obviamente deben ser iguales, dado que se calculan a partir de las sumas en la parte baja de las dos subtablas.) En realidad, tales varianzas no representan un interés central —únicamente se desea remover la varianza debida a los reactivos, de la varianza total—. El interés central reside en las varianzas individuales y en las varianzas del error, que se encuentran encerradas por un círculo en las subtablas. La varianza total de las ecuaciones 27.3, 27.4 y 27.5 es interesante, ya que es un índice de las diferencias entre individuos. Es una medida de las diferencias individuales. En lugar de escribir V_t, entonces se escribe V_{ind}, lo cual significa la varianza resultante de las diferencias individuales. Al utilizar (27.4) o (27.5) se obtienen coeficientes de confiabilidad de .92 para los datos de I, y de .45 para los datos de II. Los datos hipotéticos de I son confiables; los de II no lo son en la misma medida.

Con la ecuación 27.4:

$$r_{tt} = 1 - \frac{V_e}{V_{ind}} = 1 - \frac{.81}{10.08} = .92 \quad r_{tt} = 1 - \frac{2.60}{4.70} = .45$$

Con la ecuación 27.5:

$$r_{tt} = \frac{V_{ind} - V_e}{V_{ind}} = \frac{10.08 - 0.81}{10.08} = .92 \quad r_{tt} = \frac{4.70 - 2.60}{4.70} = .45$$

Impares-pares:

$$r_{tt} = .91 \quad r_{tt} = .32$$

Quizás la mejor forma para entender lo anterior sea regresar a la ecuación 27.3. Ahora se escribe $r_{tt} = V_\infty / V_{md}$. Si se tuviera un camino directo para calcular V_∞, se podría calcular rápidamente r_{tt}, pero como se vio antes, no existe un camino directo. Sin embargo, existe una forma para estimarlo. Si se encuentra una forma para estimar V_e, la varianza de error, entonces el problema está resuelto debido a que V_e puede restarse de V_{md} para producir un estimado de V_∞. En efecto, es posible ignorar V_∞ y restar la proporción V_e / V_{md} de 1 y obtener r_{tt}. Ésta es una forma perfectamente aceptable para calcular r_{tt} y para conceptualizar la confiabilidad. La lógica de $V_{md} - V_e$ tal vez sea más fructífera y se ligue mejor con la discusión previa sobre los componentes de la varianza.

En el capítulo 13 se estableció que cada problema estadístico tiene una cantidad total de varianza, y que cada fuente de varianza contribuye a esta varianza total. Ahora se tradu-

TABLA 27.2 *Demostración de confiabilidad y cálculo de los coeficientes de confiabilidad (ejemplos hipotéticos)*

I: $r_{tt} = .92$

Individuos	Reactivos *a*	*b*	*c*	*d*	Σ_t	$\Sigma_{impares}$	Σ_{pares}
1	6	6	5	4	21	11	10
2	4	6	5	3	18	9	9
3	4	4	4	2	14	8	6
4	3	1	4	2	10	7	3
5	1	2	1	1	5	2	3
Σ_{tr}	18	19	19	12	$\Sigma X_t = 68$ $(\Sigma X_t)^2 = 4\,624$ $\Sigma X_t^2 = 288$		

$$C = \frac{(68)^2}{20} = 231.20$$

Total = 288 − 231.20 = 56.80

$$\text{Entre reactivos} = \frac{1\,190}{5} - 231.20 = 6.80$$

$$\text{Entre individuos} = \frac{1\,086}{4} - 231.2 = 40.30$$

Fuente	*gl*	*sc*	*cm*	*F*
Reactivos	3	6.80	2.27	2.80 (n.s.)
Individuos	4	40.30	(10.08)	12.44 (.001)
Residual	12	9.70	(0.81)	
Total	19	56.80		

II: $r_{tt} = .45$

Individuos	Reactivos *a*	*b*	*c*	*d*	Σ_t	$\Sigma_{impares}$	Σ_{pares}
1	6	4	5	1	16	11	5
2	4	1	5	4	14	9	5
3	4	6	4	2	16	8	8
4	3	6	4	3	16	7	9
5	1	2	1	2	6	2	4
Σ_{tr}	18	19	19	12	$\Sigma X_t = 68$ $(\Sigma X_t)^2 = 4\,624$ $\Sigma X_t^2 = 288$		

$$C = \frac{(68)^2}{20} = 231.20$$

Total = 56.80

Entre reactivos = 6.80

$$\text{Entre individuos} = \frac{1\,000}{4} - 231.20 = 18.80$$

Fuente	*gl*	*sc*	*cm*	*F*
Reactivos	3	6.80	2.27	1 (n.s.)
Individuos	4	18.80	(4.70)	1.81 (n.s.)
Residual	12	31.20	(2.60)	
Total	19	56.80		

cirá el razonamiento del capítulo 13 al problema presente. En muestras aleatorias de la misma población, V_e y V_m deben ser iguales estadísticamente. Pero si V_m, la varianza entre grupos, es significativamente mayor que V_d, la varianza dentro de grupos (error), entonces existe algo en V_e más allá y por encima del azar. Esto es, V_e incluye la varianza de V_d y, además, un poco de varianza sistemática.

De forma similar, puede decirse que si V_{ind} es significativamente mayor que V_e, entonces existe algo en V_{ind} más allá y por encima de la varianza del error. Dicho exceso de varianza parecería que se debe a diferencias individuales en aquello que se esté midiendo. La medición apunta hacia las puntuaciones "verdaderas" de los individuos. Cuando se dice que la confiabilidad es la precisión de un instrumento de medición, se quiere indicar que un instrumento confiable de medición más o menos mide las puntuaciones "verdaderas" de individuos, siendo que el "más o menos" depende de la confiabilidad del instrumento. El hecho de que se midan las puntuaciones "verdaderas" puede inferirse únicamente a partir de las *diferencias* "verdaderas" entre individuos; aunque ninguna de ellas pueda, por supuesto, medirse de forma directa. Lo que se hace es inferir las diferencias "verdaderas" a partir de las diferencias empíricas y falibles medidas, las cuales están siempre, en cierta medida, corruptas por errores de medición.

Ahora, si existe alguna manera de eliminar de V_{ind} el efecto de los errores de medición, alguna manera de liberar a V_{ind} del error, entonces el problema se resuelve con facilidad. Tan sólo se resta V_e de V_{ind} para obtener un estimado de V_m. Entonces la proporción de la varianza "pura" respecto de toda la varianza, "pura" e "impura", es el estimado de la confiabilidad del instrumento de medición. Para resumirlo simbólicamente:

$$r_{tt} = \frac{V_m}{V_{ind}} = \frac{V_{ind} - V_e}{V_{ind}} = 1 - \frac{V_e}{V_{ind}}$$

Los cálculos reales se presentan en la parte final de la tabla 27.2.

Regresando a los datos de la tabla 27.2, analice si es posible "observar" la confiabilidad de I y la no-confiabilidad de II. Observe primero las columnas donde están registrados los totales de los individuos (Σ_i). Note que las sumas de I tienen un mayor rango que las de II: 21 – 5 = 16 y 16 – 6 = 10. Dados los mismos individuos, a mayor confiabilidad de una medida, mayor será el rango de las sumas de los individuos. Piense en el extremo: un instrumento completamente no confiable produciría sumas parecidas a las sumas producidas por números aleatorios y, por supuesto, la confiabilidad de los números aleatorios es aproximadamente de cero. (La razón F no significativa para individuos, 1.81 en II, indica que r_{tt} = .45 no es estadísticamente significativo.)

Ahora examine los órdenes de rango de los valores bajo los reactivos *a*, *b*, *c* y *d*. En I los cuatro órdenes de rango son casi iguales. Aparentemente cada reactivo de la escala de actitud está midiendo la misma cuestión. Dependiendo del grado en que los reactivos individuales produzcan los mismos órdenes de rango de individuos, la prueba será confiable. Los reactivos permanecen unidos, por decirlo así; son consistentes internamente. Note también que los órdenes de rango de los reactivos de I son casi los mismos que los órdenes de rango de las sumas.

Los órdenes de rango de los valores de los reactivos de II son bastante diferentes. Los órdenes de rango de *a* y *c* concuerdan bastante; son iguales a los de I. Sin embargo, los órdenes de rango de *a* y *b*, *a* y *d*, *b* y *d*, *c* y *d*, no concuerdan muy bien. O los reactivos están midiendo cuestiones diferentes, o no están midiendo de forma muy consistente. Esta falta de congruencia de los órdenes de rango se refleja en los totales de los individuos. A pesar de que los órdenes de rango de los totales es similar a los órdenes de rango de los totales de

I, el rango o varianza es considerablemente menor, y existe una falta de dispersión entre las sumas (por ejemplo, los tres números 16).

Se concluye la consideración de estos dos ejemplos examinando ciertas cifras en la tabla 27.2, que no fueron consideradas anteriormente. En el lado derecho de I y II se presentan las sumas de los reactivos impares ($\Sigma_{impares}$) y las sumas de los reactivos pares (Σ_{pares}). Tan sólo se suman los valores de los reactivos impares a través de los renglones: $a + c$: 6 + 5 = 11, 4 + 5 = 9, 4 + 4 = 8, etcétera, en I. Después se suman también los valores de los reactivos pares en I: $b + d$: 6 + 4 = 10, 6 + 3 = 9, etcétera. Si hubiera más reactivos, por ejemplo, a, b, c, d, e, f, g, entonces se sumarían: $a + c + e + g$ para las sumas impares, y $b + d + f$ para las sumas pares. Para calcular el coeficiente de confiabilidad, se calcula la correlación producto-momento entre las sumas impares y las sumas pares, y después se corrige el coeficiente resultante con la fórmula de Spearman-Brown. Tanto las sumas de los reactivos impares como de los pares son, por supuesto, las sumas de únicamente la mitad de los reactivos de una prueba. Por ende, son menos confiables que las sumas de todos los reactivos. La fórmula de Spearman-Brown corrige el coeficiente impar-par (y otros coeficientes partidos) para el menor número de reactivos utilizados en el cálculo del coeficiente. (Se explicará más sobre esto en una sección posterior de este capítulo. También se pueden consultar varias pruebas buenas y libros de medición tales como el de Anastasi y Urbina, 1997; Brown, 1983; Friedenberg, 1995; o Sax, 1997.) Los r_n impar-par para I y II son .91 y .32, respectivamente; bastante cerca de los resultados del análisis de varianza de .92 y .45. (Con más participantes y más reactivos, los estimados generalmente son más cercanos.)

Esta simple operación quizá parezca desconcertante. Para observar que ésta es una variación del mismo tema sobre la varianza y el orden de rango, observe primero el orden de rango de las sumas de los dos ejemplos. Los órdenes de rango de Σ_{impar} y Σ_{par} son casi iguales en I, pero bastante diferentes en II. La lógica es la misma que antes. Evidentemente, los reactivos están midiendo la misma cuestión en I, pero en II los dos conjuntos de reactivos no son consistentes. Para reconstruir la discusión sobre la varianza, recuerde que al sumar la suma de los reactivos impares con la suma de los reactivos pares de cada persona, se obtiene la suma total o $\Sigma_{impares} + \Sigma_{pares} = \Sigma_t$.

Interpretación del coeficiente de confiabilidad

Si r, el coeficiente de correlación, se eleva al cuadrado, se convierte en un coeficiente de determinación. Éste brinda la proporción o porcentaje de la varianza compartida por dos variables. Si $r = .90$, entonces las dos variables comparten $(.90)^2 = 81\%$ de la varianza total de las dos variables en común. El coeficiente de confiabilidad es también un coeficiente de determinación. Teóricamente indica cuánta varianza, de la varianza total de una variable medida, es "verdadera". Si se tuvieran las puntuaciones "verdaderas" y se pudieran correlacionar con las puntuaciones de la variable medida, y se elevara al cuadrado el coeficiente de correlación resultante, entonces se obtendría el coeficiente de confiabilidad.

Una representación simbólica servirá para aclarar esto. Sea $r_{t\infty}$ el coeficiente de correlación entre las puntuaciones obtenidas y las puntuaciones "verdaderas", X_∞. El coeficiente de confiabilidad se define de la siguiente manera:

$$r_{tt} = (r_{t\infty})^2 \qquad (27.6)$$

Aunque no es posible calcular $r_{t\infty}$ de forma directa, es útil entender la lógica del coeficiente de confiabilidad en dichos términos teóricos. La correlación de la puntuación verdadera con la puntuación observada con frecuencia se conoce como el *índice de confiabilidad*.

Puesto que una puntuación verdadera es algo que existe pero que no puede medirse, es obvio que el índice de confiabilidad no puede calcularse directamente. Como resultado, el coeficiente de confiabilidad no puede obtenerse de manera directa, por lo menos a través de este método. No obstante, existen varias formas para calcular la confiabilidad de las mediciones. Magnusson (1967) se refiere a ellas como métodos prácticos para estimar la confiabilidad. El primero consiste en aplicar el mismo instrumento de medición al mismo grupo de personas, en dos ocasiones diferentes. El lapso de tiempo entre las dos ocasiones depende del tipo y del propósito de las mediciones. Por lo común, se elige un intervalo de tiempo entre ambas aplicaciones, para que haya suficiente disminución del recuerdo sobre las respuestas. La realización adecuada del procedimiento conduce a dos mediciones por persona, las cuales, dadas en pares, se utilizan en una fórmula para calcular la correlación. Dicha correlación entre las puntuaciones de la ocasión 1 y de la ocasión 2 se denomina *confiabilidad test-retest*. Sirve para medir la estabilidad a través del tiempo. Ésta no es una buena manera para calcular el coeficiente de confiabilidad si el abandono escolar es alto o si los organismos que se están midiendo pasarán por un cambio drástico en el desarrollo, entre el periodo 1 y el periodo 2. Si el instrumento de medición es una prueba de vocabulario, la confiabilidad test-retest puede no resultar fructífera si la prueba se aplica, en dos o más ocasiones, a niños que están expuestos a un ambiente educativo donde su vocabulario se desarrolla rápidamente. Otra interpretación teórica es considerar que cada X_∞ puede ser la media de un gran número de X_n derivadas de la aplicación de una prueba a un individuo un gran número de veces, si lo demás permanece igual. La idea que subyace a esto se explicó anteriormente. La primera aplicación de la prueba produce, por ejemplo, un cierto orden de rango de los individuos. Si la segunda, tercera o más mediciones tienden a producir aproximadamente el mismo orden de rango, entonces la prueba es confiable, lo cual representa una interpretación de estabilidad o test-retest de la confiabilidad.

Otro método que puede utilizarse para calcular el coeficiente de confiabilidad consiste en desarrollar dos *formas equivalentes* o *paralelas* del instrumento de medición. En términos de prueba, esto implicaría crear dos formas de la prueba. Las dos formas serían equivalentes, pero no idénticas. Estarían compuestas de reactivos similares, posiblemente del mismo banco de reactivos. Cada persona estaría sujeta a mediciones por medio de los dos instrumentos. Como resultado, cada persona tendría, entonces, dos puntuaciones y, nuevamente, los pares de puntuaciones serían utilizados en una fórmula de correlación para calcular la correlación. Tal correlación sería considerada como una forma paralela o equivalente de confiabilidad. Dicho método posee la ventaja de minimizar las deserciones escolares. Además, tampoco hay que preocuparse demasiado respecto a si las personas que se están midiendo recordarán las respuestas. Sin embargo, las formas paralelas tienen algunos problemas. Por un lado, se requiere que el investigador realice dos formas de la prueba, las que necesitarían tener medias y desviaciones estándar que sean equivalentes estadísticamente. También, el procedimiento deseable requeriría que las personas que se miden tengan que estar sujetas a mediciones durante un periodo más largo y por ende serían susceptibles a la fatiga y el aburrimiento. Si esto sucede, entonces se afectaría su desempeño en los últimos reactivos, lo que podría contribuir a disminuir el coeficiente de confiabilidad.

La tercera categoría para calcular el coeficiente de confiabilidad se denomina *consistencia interna*. Existen varios métodos para obtener la consistencia interna. Cada método depende de ciertos supuestos que pueden hacerse sobre las mediciones. El primero se llama *confiabilidad por mitades*; el segundo, *coeficiente alfa*, y el tercero, *fórmulas 20 y 21 de Kuder-Richardson* (KR-20, KR-21). Aunque en el siguiente análisis se utilizará el término *prueba* para designar al instrumento de medición, no necesariamente tiene que ser una prueba en sí. Como brevemente se mencionó y demostró antes, la confiabilidad por mita-

des implica dividir la prueba en dos mitades. El objetivo es obtener dos mitades iguales o equivalentes, lo cual se logra sumando todas las respuestas a los reactivos de la primera mitad, o sumando todas las respuestas a los reactivos de la segunda mitad. Si todos los reactivos son homogéneos, entonces las dos mitades serán iguales. Si la prueba inicia con los reactivos más fáciles y progresa hacia los más difíciles, entonces el método mencionado previamente no será efectivo en producir mitades iguales. El método recomendado aquí sería sumar todas las respuestas a los reactivos impares para crear un total, y luego sumar todas las respuestas a los reactivos pares para crear el otro total. En cualquiera de los casos anteriores, cada persona tendrá dos puntuaciones de mitad de suma. Estas puntuaciones se correlacionan utilizando la fórmula estándar. La correlación resultante se nombraría "confiabilidad por mitades". Como se demostró en Magnusson (1967), Allen y Yen (1979) y en el trabajo clásico de Gullikson (1950) con reactivos homogéneos, a mayor tamaño de la prueba (más reactivos), habrá mayor confiabilidad; a menor tamaño de la prueba (menos reactivos), habrá menor confiabilidad. Con el método de confiabilidad por mitades, ya no se está hablando acerca de una confiabilidad de la prueba completa: la confiabilidad por mitades subestimará la confiabilidad real, pues ahora se trata de la correlación de dos mitades de la prueba. Al utilizar la confiabilidad por mitades se necesita utilizar una de tres fórmulas para estimar la confiabilidad de la prueba completa, basado en valores de la mitad de ella.

Una de estas fórmulas es la fórmula profética de Spearman-Brown, la cual tiene otros usos además de la estrategia por mitades. Con el uso de esta fórmula, junto con el supuesto de que las mitades son iguales, puede calcularse un estimado de la confiabilidad de la prueba completa. La fórmula de Spearman-Brown es:

$$r_{tt} = \frac{nr_{tt}}{1 + (n - 1)r_{tt}}$$

Para la estrategia por mitades, n se establece igual a 2. La r_{tt} es la confiabilidad por mitades, y la r_{tt} es la confiabilidad estimada para la prueba completa.

Las otras dos fórmulas son distintas en apariencia, pero ambas tienen el mismo propósito. Antes de describirlas, es necesario reiterar que la fórmula de Spearman-Brown puede aplicarse a otras situaciones de confiabilidad (véase Anastasi y Urbina, 1997). También podría emplearse cuando el investigador esté relativamente seguro de que las dos mitades son iguales. Si existe cualquier duda respecto a la homogeneidad de las mitades, no debe utilizarse la fórmula Spearman-Brown, ya que sobrestimará la confiabilidad de la prueba completa. En su lugar, es preferible utilizar la fórmula de Rulon o la fórmula de Guttman (Magnusson, 1967). Ambas toman en cuenta las diferencias entre las mitades. Tanto la fórmula de Rulon como la fórmula de Guttman estiman la confiabilidad de la prueba completa sin el uso de la confiabilidad por mitades.

La fórmula de Rulon es

$$r_{tt} = 1 - \frac{V_d}{V_t} = 1 - \frac{V_{(a-b)}}{V_t}$$

y la fórmula de Guttman es

$$r_{tt} = 2\left[1 - \frac{(V_a + V_b)}{V_t}\right]$$

donde a representa el total de la primera mitad de puntuaciones; y b, el total de la segunda mitad de puntuaciones. V_d es la varianza de la diferencia de las puntuaciones ($d = a - b$), V_t es la varianza de las puntuaciones totales ($t = a + b$). V_a es la varianza del total de la primera mitad de puntuaciones; y V_b, la varianza del total de la otra mitad de puntuaciones.

Para sintetizar, los reactivos de la prueba se consideran homogéneos. Esta interpretación, en efecto, se reduce a la misma idea de otras interpretaciones: precisión. Tome cualquier muestra aleatoria de reactivos de la prueba y cualquier otra muestra aleatoria diferente de reactivos de la misma. Trate cada muestra como una subprueba separada. Entonces, cada individuo tendrá dos puntuaciones: una X_a para una submuestra, y otra X_t para la otra submuestra. Se correlacionan los dos conjuntos, y se continúa el proceso indefinidamente. La intercorrelación promedio de las submuestras (correlacionadas por medio de la fórmula Spearman-Brown) demuestra la consistencia interna de la prueba. Pero esto significa realmente que cada submuestra —si la prueba es confiable— tiene éxito en producir aproximadamente el mismo orden de rango de los individuos. Si no es así, entonces la prueba no es confiable.

La confiabilidad por mitades está basada en dos mitades que generalmente se consideran equivalentes o paralelas. Si este concepto se lleva más allá al considerar cada reactivo como una prueba paralela separada, es posible derivar algunas de las medidas de confiabilidad que se encuentran comúnmente en la literatura sobre investigación psicológica y educativa. En 1937, Kuder y Richardson desarrollaron esta idea, la cual resultó en dos de las fórmulas de confiabilidad más utilizadas para la consistencia interna: KR-20 y KR-21. Están numeradas de esta forma a causa de que la KR-20 fue la vigésima ecuación en su artículo, y la KR-21 fue la vigésimo primera ecuación. Ambas asumen que cada reactivo tiene la misma media y la misma varianza. Las fórmulas de Kuder-Richardson son aplicables a instrumentos de medición (por ejemplo, pruebas) con un sistema dicotómico o binario de calificación de respuesta. Un ejemplo de calificación dicotómica son los reactivos que se califican como correctos (1) o incorrectos (0). Las pruebas con respuestas de verdadero-falso también se consideran como un sistema dicotómico de calificación. Si se elige que p sea la proporción de receptores de la prueba que responden correctamente el reactivo i (o que se considera "verdadero"), entonces q_i es la proporción que responde incorrectamente el reactivo i (o que se considera "falso"). k es el número de reactivos en la prueba. Con esta información, la fórmula KR-20 se ve así:

$$r_{tt} = \frac{k}{k-1}\left(\frac{V_t - \sum p_i q_i}{V_t}\right)$$

Si se asume que cada reactivo tiene las mismas p_i y q_i, entonces $\sum p_i q_i$ puede reemplazarse por $kp_i q_i$. Al hacer esto se llega a KR-21.

$$r_{tt} = \frac{k}{k-1}\left(\frac{V_t - kp_i q_i}{V_t}\right)$$

la cual puede simplificarse aún más a:

$$r_{tt} = \frac{k}{k-1}\left(1 - \frac{Mk - M^2}{kV_t}\right)$$

donde k es el número de reactivos y M es la media del total de las puntuaciones. En esencia KR-21 es un caso especial de KR-20, donde $p_i q_i$ (también conocido como dificultades o

respaldo de los reactivos) son iguales. Si un investigador desea obtener el estimado de confiabilidad más conservador, para un instrumento con reactivos que usan calificación binaria, entonces se recomienda esta fórmula. Observe que este coeficiente subestimará KR-20 si las dificultades o respaldo de los reactivos tienen un rango amplio.

A manera de recordatorio, las fórmulas KR-20 y KR-21 son aplicables cuando los reactivos de un instrumento de medición (por ejemplo, una prueba) tienen calificación binaria o la escala de respuestas es dicótoma. Si el formato de calificación o de respuesta no es binario, esta fórmula no puede utilizarse. En el periodo entre el desarrollo de Kuder-Richardson en 1937, y el desarrollo del coeficiente alfa de Cronbach en 1951, se desarrollaron muchas pruebas psicológicas con base en un sistema binario de respuesta. Con la creación de Cronbach (1951), los investigadores fueron capaces de evaluar la confiabilidad de consistencia interna de su instrumento, el cual tenía diferentes escalas de calificación y de respuesta. De hecho, a través de una prueba matemática es posible demostrar que las fórmulas de Kuder-Richardson son casos especiales del coeficiente alfa de Cronbach o alfa de Cronbach. De este rango de coeficientes de confiabilidad, el coeficiente alfa es el más general. Con éste ahora es posible que un investigador encuentre la confiabilidad de instrumentos que utilicen escalas de Likert. La fórmula del alfa de Cronbach es la siguiente:

$$r_{tt} = \alpha = \frac{k}{k-1}\left(1 - \frac{\sum V_i}{V_t}\right)$$

Un método alternativo para escribir el coeficiente alfa, utilizando la intercorrelación entre reactivos, es

$$r_{tt} = \frac{n\bar{r}_{tt}}{1 + (n-1)\bar{r}_{tt}}$$

donde $\bar{r}_{tt}$ es la media de las correlaciones inter-reactivos. Lo que esto significa, esencialmente, es que si se correlacionara cada reactivo con cada uno de los demás reactivos del instrumento, se encontrara la media de dichas correlaciones y después se insertara la media de las correlaciones inter-reactivo en la fórmula de Spearman-Brown, entonces se obtendría el coeficiente alfa o la fórmula de Kuder-Richardson.

Cabe señalar que el ejemplo computacional realizado anteriormente en este capítulo constituye un ejemplo donde se puede utilizar el análisis de varianza para determinar el coeficiente de confiabilidad, y debe ser equivalente al coeficiente alfa.

El error estándar de la media y el error estándar de medición

Dos aspectos importantes de la confiabilidad son la confiabilidad de las medias y la confiabilidad de las medidas individuales, los cuales se relacionan con el error estándar de la media y el error estándar de la medición. En estudios de investigación, generalmente el error estándar de la media y de estadísticos relacionados —como el error estándar de las diferencias entre medias y el error estándar de un coeficiente de correlación— es el más importante de ellos. Puesto que el error estándar de la media se discutió de manera considerable en un capítulo anterior, sólo es necesario decir aquí que la confiabilidad de estadísticos específicos es otro aspecto del problema general de confiabilidad. El error

TABLA 27.3 *Confiabilidad y error estándar de medición (ejemplo hipotético)*

	X_t	X_∞	X_e
	2	1	1
	1	2	−1
	3	3	0
	3	4	−1
	6	5	1
Σ:	15	15	0
M:	3	3	0
V:	2.8	2.0	.80

$$r_{tt} = 1 - \frac{V_e}{V_t} = 1 - \frac{V_e}{V_O} = 1 - \frac{.80}{2.80} = 0.71 \qquad r_{\infty\infty} = 0.845$$

$$r_{tt} = \frac{V_\infty}{V_t} = \frac{2.00}{2.80} = 0.71 \qquad r_{tt} = r^2_{\infty\infty} = (.845)^2 = 0.71$$

$$VE_{med} = V_t(1 - r_{tt}) = 2.80\,(1 - 0.71) = 0.81$$

$$EE_{med} = DE_t\sqrt{1 - r_{tt}} = \sqrt{VE_{med}}\,\sqrt{0.81} = 0.90$$

estándar de medición, o su cuadrado, la varianza estándar de medición, necesita definirse e identificarse, aunque sea de manera breve. Esto se hará mediante un ejemplo simple.

Un investigador mide las actitudes de cinco individuos y obtiene las puntuaciones presentadas bajo la columna llamada X_t en la tabla 27.3. Suponga, además, que las puntuaciones "verdaderas" de actitud de los cinco individuos son aquellas presentadas bajo la columna llamada X_∞. (Sin embargo, recuerde que en la realidad nunca es posible conocer estas puntuaciones.) Puede notarse que el instrumento es confiable. A pesar de que sólo una de las puntuaciones obtenidas es exactamente igual a su puntuación acompañante "verdadera", las diferencias, entre las puntuaciones obtenidas que son diferentes y las puntuaciones "verdaderas", son pequeñas. Tales diferencias se presentan bajo la columna llamada "X_e": son "puntuaciones de error". Evidentemente el instrumento es bastante preciso. El cálculo de r_{tt} confirma dicha impresión: .71.

Una medida muy directa de la confiabilidad del instrumento puede obtenerse al calcular la varianza o la desviación estándar o las puntuaciones de error (X_e). La varianza de las puntuaciones de error y las varianzas de las puntuaciones X_t y X_∞ se calcularon y se incluyeron en la tabla 27.3. La varianza de las puntuaciones de error ahora se nombran, justificadamente, como *varianza estándar de medición*, la cual podría llamarse con mayor precisión "varianza estándar de los errores de medición". La raíz cuadrada de dicho estadístico se denomina *error estándar de medición*. La varianza estándar de medición se define de la siguiente manera:

$$VE_{med} = V_t\,(1 - r_{tt}) \tag{27.7}$$

En efecto, sólo es posible calcular tal estadístico, si se conoce el coeficiente de confiabilidad. Note que si existe alguna forma para estimar VE_{med}, entonces es posible calcular el coeficiente de confiabilidad. Esto requiere de mayor investigación.

Se inicia con la definición de confiabilidad dada anteriormente: $r_{tt} = V_\infty / V_t = 1 - V_e / V_t$. Una ligera manipulación algebraica produce la varianza estándar de medición:

$$r_{tt} = 1 - \frac{V_e}{V_t}$$

$$r_{tt} V_t = V_t - V_e$$

$$V_e = V_t - r_{tt} V_t$$

$$V_e = V_t(1 - r_{tt})$$

La parte derecha de la ecuación es igual a la parte derecha de la ecuación 27.7. Por lo tanto, $V_e = VE_{med}$, o la varianza de error utilizada anteriormente en el análisis de varianza es la varianza estándar de medición. La varianza estándar de medición y el error estándar de medición del ejemplo se calcularon en la tabla 27.3, y son .81 y .90, respectivamente. Como muestran los libros de texto sobre medición (por ejemplo, Anastasi y Urbina, 1997), sirven para interpretar puntuaciones individuales de pruebas. Dicha interpretación no será discutida aquí; tales estadísticos se han incluido sólo para demostrar la conexión entre la teoría original y las formas para determinar la confiabilidad.

Otro cálculo de la tabla 27.3 requiere de una explicación. Si se correlacionan las puntuaciones X_t y X_∞, se obtiene un coeficiente de correlación de .845. Ahora se obtiene este coeficiente $r_{t\infty}$ de forma directa, y se eleva al cuadrado para obtener el coeficiente de confiabilidad (ecuación 27.6). Este último es, por supuesto, igual al anterior: .71.

Incremento de la confiabilidad

El principio que subyace al incremento de la confiabilidad es el llamado anteriormente principio *maxmincon*, en una forma ligeramente diferente: "Maximizar la varianza de las diferencias individuales y minimizar la varianza del error." La ecuación 27.4 indica con claridad tal principio. A continuación se describe el procedimiento general.

Primero, se escriben sin ambigüedades los reactivos de los instrumentos de medición psicológica o educativa. Un evento ambiguo llega a interpretarse en más de una forma. Un reactivo ambiguo permite que la varianza del error se introduzca silenciosamente, debido a que los individuos pueden interpretar el reactivo de forma diferente. Dichas interpretaciones tienden a ser aleatorias y, por lo tanto, incrementan la varianza del error y disminuyen la confiabilidad.

Segundo, si un instrumento no es lo suficientemente confiable, deben añadirse más reactivos del mismo tipo y calidad, por lo común, aunque no necesariamente, incrementará la confiabilidad en una cantidad predecible. El añadir más reactivos incrementa la posibilidad de que la X_t de cualquier individuo esté cerca de su X_∞. Ello es una cuestión del muestreo de la propiedad del espacio o del reactivo. Con pocos reactivos, puede surgir un error grande por el azar. Con más reactivos puede no ser tan grande. La probabilidad de que se balancee por otro error aleatorio en sentido inverso es mayor cuando hay más reactivos. En síntesis, una mayor cantidad de reactivos incrementa la probabilidad de una medición precisa. (Recuerde que cada X_t es la suma de los valores de los reactivos, para cada individuo.)

En tercer lugar, la especificación de instrucciones claras y estándar tiende a reducir los errores de medición. Siempre se debe tener mucho cuidado al escribir las instrucciones para expresarlas con claridad, ya que las instrucciones ambiguas incrementan la varianza del error. Además, los instrumentos de medición deben aplicarse siempre bajo condiciones estándar, bien controladas y similares. Si las situaciones de aplicación difieren, de nuevo puede introducirse varianza del error. En los campos de la psicología y educación,

una prueba que tiene uniformidad de aplicación y calificación se denomina *prueba estandarizada*. Por lo tanto, las pruebas estandarizadas son aquellas que se han sujetado al rigor de la reducción de la varianza del error.

¿Entonces cómo saber si se han escrito reactivos ambiguos o claros? ¿Cómo saber si los reactivos añadidos para intentar incrementar la confiabilidad son del mismo tipo y calidad? Existe un conjunto de procedimientos estadísticos llamados *análisis de reactivos*, que ayudan a responder tales preguntas. El análisis de reactivos se utiliza para incrementar tanto la confiabilidad como la validez de una prueba, lo cual se logra al evaluar cada reactivo de forma separada para determinar si el reactivo es bueno o pobre. Si el reactivo mide o no lo que se desea que mida es cuestión de validez. La validez se analiza en el capítulo 28. En pruebas donde las respuestas se evalúan como correctas e incorrectas (como las pruebas cognitivas), los reactivos se evalúan en términos de su nivel de dificultad. En pruebas donde no hay respuestas correctas o incorrectas (como las que se encuentran en pruebas afectivas), se utilizaría el índice de acuerdos en lugar de la dificultad. El índice de dificultad es una razón simple del número de personas que responden correctamente el reactivo y el número total de personas que toman la prueba. El índice de acuerdos se calcula como la razón del número de personas que selecciona una respuesta, entre el número total de personas que responden la prueba. Por lo tanto, en esencia, el índice de dificultad y el índice de acuerdos son similares en su cálculo.

$$\text{Dificultad del reactivo} = \frac{\textit{número de personas que responden correctamente un reactivo}}{\textit{número total de personas que toma la prueba}}$$

$$\text{Índice de acuerdos} = \frac{\textit{número de personas que selecciona una respuesta}}{\textit{número total de personas que toma la prueba}}$$

Para el índice de dificultad, a mayor valor, más fácil será el reactivo. Lo anterior indica que más personas respondieron correctamente el reactivo. Reactivos con índices de 0.0 o 1.00 contribuyen muy poco a la prueba, en términos de la información que brindan acerca de las diferencias entre las personas. Cuando cada estudiante responde correctamente casi todos los reactivos en una prueba fácil de matemáticas, esto revela muy poco acerca de la diferencia de las personas en habilidades matemáticas. Por otro lado, una prueba que consista de reactivos demasiado difíciles tampoco revela qué tanto difieren los individuos. No importa cuáles sean sus habilidades, todos los individuos responderán de forma incorrecta esos reactivos. Por regla general, la mayoría de los creadores de pruebas concuerdan en que los mejores reactivos, en términos de dificultad y de acuerdo, son aquellos con valores entre .5 y .7. Algunos recomiendan combinar reactivos de diferentes niveles de dificultad, pero que tengan un índice general entre .5 y .7.

Después de la dificultad y del acuerdo, el siguiente índice para el análisis de reactivos es el *índice de discriminación de reactivos*. Dicho estadístico es el que indicará al investigador (en pruebas cognitivas) qué tan efectivamente el reactivo fue capaz de discriminar entre puntuaciones altas y puntuaciones bajas. Se considera un buen reactivo a aquel que es contestado correctamente por las personas con alta puntuación, y contestado erróneamente por aquellos con baja puntuación. Cuando así sucede, el reactivo tiene la discriminación máxima. El índice de discriminación de reactivos funciona mejor para pruebas cognitivas, las cuales son pruebas que tienen respuestas correctas e incorrectas. En pruebas de tipo afectivo (por ejemplo, de personalidad), donde no hay respuestas correctas e

incorrectas, se utiliza la correlación de la puntuación del reactivo con la puntuación total, aunque ésta también puede utilizarse con pruebas cognitivas.

Con el índice de discriminación de los reactivos, el investigador primero determina el grupo con puntuación más alta y el grupo con puntuación más baja. Para hacerlo se utilizan las puntuaciones totales. Es recomendable que los dos grupos sean iguales en términos del número de personas; éste varía dependiendo del número de personas que tomó la prueba. Después se cuenta el número de personas, dentro de cada grupo, que respondieron correctamente el reactivo. Se calcula una puntuación de diferencia entre el número de personas en el grupo de alta puntuación, que respondieron correctamente el reactivo, y el número de personas del grupo de baja puntuación que respondieron correctamente el mismo reactivo. El índice de discriminación del reactivo es la razón de la diferencia y el número de personas en el grupo de alta puntuación. Se podría haber utilizado como denominador de este cálculo el número de personas del grupo de baja puntuación; pero el número debe ser el mismo:

$$\text{Índice de discriminación del reactivo } i = \frac{P_A - P_B}{\text{\# de personas en el grupo de alta puntuación}}$$

donde P_A es el número de personas en el grupo de alta puntuación que respondieron correctamente el reactivo, y P_B es el número de personas del grupo de baja puntuación que respondieron correctamente el mismo reactivo.

Valores de 0.0, 1.0 y –1.0 son raros. Si el índice es negativo, el reactivo posee discriminación invertida. Esto indicaría al investigador que algo anda definitivamente mal con este reactivo. Se espera que los reactivos tengan valores positivos; a mayor valor, mayor discriminación.

En el caso de la correlación del reactivo con la puntuación total, el investigador, en esencia, correlacionaría la puntuación de cada reactivo o respuesta con la puntuación total. La idea aquí es que si el reactivo es parte de un todo —un todo que mide algo que se desea— debe tener una alto valor de correlación con el total. Recuerde, puesto que se espera que los reactivos sean homogéneos, la correlación de cada reactivo con la puntuación total debe ser alta. Un reactivo que tiene una baja correlación con el total se interpreta como un reactivo que está midiendo algo que difiere de aquello que los demás reactivos están midiendo. El reactivo no es homogéneo con los demás reactivos. Con las computadoras de alta velocidad y la disponibilidad de programas estadísticos, un investigador obtiene dichas correlaciones muy fácilmente. Friedenberg (1995) ofrece una presentación muy buena sobre la manera de calcular tales índices.

El análisis de reactivos con el empleo de estos métodos más tradicionales funciona relativamente bien. Sin embargo, existe un nuevo desarrollo caracterizado por mejorías claras respecto a los métodos tradicionales. Este "nuevo elemento" en el análisis de reactivos se denomina *Teoría de Respuesta al Item* o *TRI*. La TRI involucra mucho más matemáticas que el método tradicional. Su meta principal consiste en clasificar la dificultad o acuerdo de los reactivos. A causa de su complejidad matemática, es mejor realizarlo por medio de programas computacionales. Una compañía llamada Assessment Systems Corporation distribuye varios de los programas a través de Lawrence Erlbaum Associates. Este método esencialmente implica el uso de la *curva característica del reactivo* (ítem) con la teoría del *rasgo latente*. En la teoría del rasgo latente se asume que el desempeño de la prueba puede ser explicado por la posición de quien toma la prueba, sobre una característica hipotética e inobservable (por ejemplo, un rasgo). No se implica que el rasgo cause el comportamiento ni que dicho rasgo exista física o fisiológicamente. Los rasgos latentes son meros constructos estadísticos creados a partir de datos empíricos. La medición básica utilizada en la TRI es

una probabilidad. Es la probabilidad de que una persona con una habilidad específica o rasgo latente responda correctamente un reactivo, con un nivel específico de dificultad. Con reactivos que no se califican como correctos e incorrectos, la TRI aun puede calcular la probabilidad de que una persona con cierta característica dé una respuesta específica, basada en los acuerdos de tal reactivo.

La curva característica del reactivo es una gráfica de la relación entre la puntuación que obtiene en la prueba la persona que la toma y el desempeño en un reactivo en particular. La puntuación de la prueba, por supuesto, mide qué cantidad del atributo o rasgo tiene el individuo. El desempeño en un reactivo en particular por lo común se expresa en forma de probabilidad o proporción. Los mejores reactivos tenderán a exhibir un patrón donde aquellos con altas puntuaciones tiendan a responder correctamente el reactivo; mientras que aquellos con puntuaciones bajas tiendan a responder incorrectamente el mismo reactivo. A mayor pendiente de la curva, de las puntuaciones bajas hacia las puntuaciones altas (pendiente positiva), mayor será el poder discriminativo de ese reactivo. Los reactivos con discriminación negativa tienen una pendiente negativa y tienen un problema que requiere mayor análisis. La curva característica del reactivo también puede ofrecer una medida de la dificultad del reactivo. Al tomar el nivel .50 de probabilidad o proporción y encontrar la puntuación total correspondiente de la prueba para ese nivel, esta puntuación total puede utilizarse como medida de la dificultad. La puntuación total de la prueba correspondería al punto donde el 50% de quienes tomaron la prueba respondieron correctamente el reactivo. Esto difiere ligeramente del índice de dificultad del reactivo que se analizó antes; pero es tan útil como él. Por medio del uso del ajuste matemático y estadístico de la curva, un investigador obtiene índices de discriminación y dificultad de las curvas características de los reactivos. El ajuste de la curva no lineal utilizado en estos procedimientos va más allá del alcance de este libro. Se refiere al lector a estupendas obras que tratan el tema: Allen y Yen (1979), Baker (1992), Crocker y Algina (1986) y Wright y Stone (1979).

El valor de la confiabilidad

Para ser interpretable, una prueba debe ser confiable. A menos que se pueda depender de los resultados de la medición de las propias variables, no es posible determinar, con alguna confianza, las relaciones entre las variables. Puesto que la medición no confiable es medición sobrecargada de error, la determinación de relaciones se convierte en una tarea difícil y poco convincente. ¿Es bajo un coeficiente de correlación obtenido entre dos variables, debido a que una o ambas medidas no sean confiables? ¿Una razón *F* del análisis de varianza es no significativa debido a que la relación hipotetizada no existe, o debido a que la medida de la variable dependiente no es confiable?

La confiabilidad, aunque no es el aspecto más importante de la medición, es bastante importante. En cierto sentido, esto es como el problema del dinero: su ausencia constituye el verdadero problema. Una confiabilidad alta no es garantía de buenos resultados científicos; pero no puede haber buenos resultados científicos sin confiabilidad. En resumen, la confiabilidad es una condición necesaria, pero no suficiente, del valor de los resultados de la investigación y su interpretación.

En este punto es necesario plantear la pregunta: ¿qué tan alto se requiere que sea el coeficiente de confiabilidad? No existe una respuesta rápida y rigurosa a esta pregunta. Por alguna razón, diversos investigadores han establecido .70 como el límite entre confiabilidades aceptables y no aceptables; sin embargo, no existe ninguna evidencia para apoyar esta regla arbitraria. De hecho, la mayoría de los autores de los libros de texto (sobre medición) no establecen dicho valor. Anastasi y Urbina (1997), por ejemplo, no

mencionan tal regla. Nunnally (1978) afirma que un nivel satisfactorio de confiabilidad depende de cómo se utilice la medida. En algunos casos un valor de confiabilidad de .50 o .60 es aceptable; mientras que en otras un valor de .90 es apenas aceptable. Un valor bajo de confiabilidad puede ser aceptable si el instrumento de medición posee una validez alta. Gronlund (1985) señala que la mayoría de las pruebas realizadas por maestros poseen confiabilidades de entre .60 y .85, y aun así son útiles en decisiones instruccionales. Gronlund también brinda consideraciones que deben tenerse al decidir si un valor de confiabilidad es aceptable. Todas las consideraciones se centran en qué tipo de decisión se toma al utilizar la prueba o el instrumento de medición. Si la decisión tomada por medio de la prueba es importante, final, irreversible, inconfirmable, concierne a individuos o tiene consecuencias duraderas, entonces es necesario un alto nivel de confiabilidad. Si la decisión tiene poca importancia, tomada en una etapa temprana, reversible, confirmable por medio de otros datos, concierne a grupos o tiene efectos temporales, entonces es aceptable un valor bajo de confiabilidad.

RESUMEN DEL CAPÍTULO

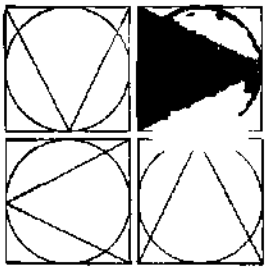
1. Este capítulo examina principalmente la teoría clásica de la confiabilidad. También contempla algunos de los desarrollos "más novedosos" en esta área.
2. La confiabilidad se define como la consistencia o estabilidad del instrumento de medición.
3. La teoría clásica de las pruebas creó la ecuación: $X_t = X_{vv} + X_e$, donde X_t es la puntuación observada, X_{vv} es la puntuación verdadera y X_e es la puntuación de error.
4. La confiabilidad y la validez se confunden a menudo debido a que ambas tratan con la precisión de las mediciones. No obstante, la confiabilidad está poco relacionada con el hecho de si el instrumento realmente mide lo que se desea. Su aspecto de precisión se refiere a la medición de la puntuación "verdadera".
5. Una medición puede ser confiable e inválida al mismo tiempo. El instrumento de medición puede medir algo de forma imprecisa todo el tiempo.
6. El índice de confiabilidad es de interés; es la correlación entre las puntuaciones verdaderas y las puntuaciones observadas. Sin embargo, las puntuaciones verdaderas no son observables.
7. El coeficiente de confiabilidad es el cuadrado del índice de confiabilidad.
8. Métodos prácticos para obtener el coeficiente de confiabilidad son:

   test-retest, formas paralelas, consistencia interna

9. La consistencia interna puede obtenerse a través de uno de los siguientes métodos: por mitades, por las fórmulas 20 y 21 de Kuder-Richardson, y por el coeficiente alfa de Cronbach.
10. El error estándar de medición indica qué cantidad de error hay en el coeficiente de confiabilidad.
11. Para incrementar la confiabilidad se pueden escribir mejores reactivos, añadir más reactivos similares y estandarizar la administración y la calificación del instrumento de medición y las respuestas.
12. El análisis de reactivos brinda información sobre qué tan buenos o qué tan pobres son los reactivos dentro del instrumento de medición.
13. Qué tan alto debe ser el coeficiente de confiabilidad, para ser aceptable, depende del tipo de decisión a tomar y de las condiciones bajo las cuales se determinó el coeficiente.

SUGERENCIAS DE ESTUDIO

1. ¿En qué difiere la teoría de la generalización de la teoría clásica de las pruebas?
2. De las siguientes, ¿cuál considera usted que es más útil para los investigadores:? ***a)*** validez, o ***b)*** confiabilidad. Justifique su elección.
3. Describa algunos de los problemas con ***a)*** la confiabilidad test-retest y ***b)*** las formas paralelas de confiabilidad. Señale un ejemplo donde usted usaría y no usaría cada una de éstas.
4. Dadas las siguientes situaciones enlistadas abajo, ¿cuál coeficiente de confiabilidad sería el más adecuado para cada una?
   a) Una prueba de mecanografía aplicada a un grupo de alumnos en un curso sobre el uso del procesador de palabras
   b) Una lista de problemas psicológicos utilizada por terapeutas
   c) Una prueba cognitiva de rendimiento
   d) Una prueba de ortografía con palabras de cuatro letras
   e) El número de actos "agresivos" de un chimpancé macho en un zoológico, durante el mismo periodo diario de 10 minutos
   f) Después de que un grupo de estudiantes completó una prueba, ésta se dividió en dos partes y se calcularon las puntuaciones separadas para cada estudiante: la correlación de las dos puntuaciones fue de 0.79
5. ¿Cuántos componentes diferentes puede encontrar, que fueran parte del término de error, en la ecuación de la teoría clásica de las pruebas: $X_t = X_\infty + X_e$?
6. Ofrezca una explicación referente a por qué una puntuación o medida "verdadera" nunca puede alcanzarse.
7. La confiabilidad por mitades de una prueba es de .70. ¿Cuál es la confiabilidad estimada de la prueba completa?
8. Si una confiabilidad test-retest de una prueba con 50 reactivos es de .65, ¿cuál sería la confiabilidad estimada si se añadieran 50 reactivos similares a la prueba?



CAPÍTULO 28

VALIDEZ

El tema de la validez es complejo, controvertido y especialmente importante en la investigación del comportamiento. Aquí, quizás más que en cualquier otra parte, se cuestiona la naturaleza de la realidad. Sin embargo, no es posible estudiar la validez sin investigar, tarde o temprano, el significado de las variables. Sin embargo, no es posible estudiar la validez sin tarde o temprano investigar sobre la naturaleza y el significado de las propias variables.

Cuando se miden ciertas propiedades físicas y atributos relativamente simples de personas, la validez no representa un gran problema. Más bien existe, con frecuencia, congruencia cercana y directa entre la naturaleza del objeto que se mide y el instrumento de medición. Por ejemplo, la longitud de un objeto puede medirse colocando palos marcados con un sistema numérico estándar (pies o metros) sobre el objeto. El peso es más indirecto, pero no difícil: un objeto ubicado en un contenedor desplaza al contenedor hacia aba-

jo. El movimiento del contenedor hacia abajo se registra sobre un índice calibrado (libras u onzas). Por lo tanto, con ciertos atributos físicos existe poca duda de aquello que se está midiendo.

Por otro lado, suponga que un científico educativo desea estudiar la relación entre inteligencia y rendimiento escolar o la relación entre autoritarismo y estilo de enseñanza. Ahora no existen reglas que utilizar, ni escalas con que medir el grado de autoritarismo, ni atributos físicos o de comportamiento claros que indiquen, sin lugar a dudas, el estilo de enseñanza. En tales casos es necesario inventar formas indirectas para medir propiedades psicológicas y educativas. Estas formas son, en ocasiones, tan indirectas que la validez de la medición y sus productos se vuelven dudosos.

Tipos de validez

La definición más común de validez se sintetiza en la pregunta: ¿estamos midiendo lo que creemos que estamos midiendo? El énfasis en esta pregunta está en lo que se mide. Por ejemplo, un maestro ha construido una prueba para medir la *comprensión* de los procedimientos científicos y ha incluido en la prueba sólo reactivos *factuales* sobre procedimientos científicos. La prueba no es válida ya que aunque quizás mida de manera confiable el *conocimiento factual* de los alumnos sobre los procedimientos científicos, no mide su *comprensión* de dichos procedimientos. En otras palabras, quizás mida bastante bien aquello que mide; pero no mide lo que el maestro en realidad intentaba medir.

Aunque la definición más común de validez fue expresada antes, debe enfatizarse de inmediato que no existe una validez única. Una prueba o escala es válida de acuerdo con el propósito científico o práctico de quien la utiliza. Los educadores pueden estar interesados en la *naturaleza* del rendimiento en matemáticas de los alumnos de preparatoria. Entonces ellos estarían interesados en *lo que* mide una prueba de rendimiento o aptitud matemática. Por ejemplo, ellos podrían querer conocer los factores involucrados en el desempeño de una prueba de matemáticas y sus contribuciones relativas a este desempeño. Por otro lado, podrían estar interesados en conocer a los alumnos que probablemente tendrán éxito y a aquellos que probablemente no lo tendrán en las matemáticas de preparatoria. Quizás tengan poco interés en *lo que* mide una prueba de aptitud matemática, y estén interesados ante todo en una *predicción* exitosa. En estos dos usos de las pruebas están implicados diferentes tipos de validez. Ahora se examinará un desarrollo extremadamente importante en la teoría de las pruebas: el análisis y el estudio de los diferentes tipos de validez. Aunque existan varios tipos, el investigador debe diseñar el estudio de validación sólo con un tipo de validez en mente. Algunos investigadores calculan todos los coeficientes de validez sólo para descubrir que cada uno adquiere un valor diferente.

La clasificación más importante de los tipos de validez es la que creó un comité conjunto de la Asociación Psicológica Americana, la Asociación Americana de Investigación Educativa y el Consejo Nacional de Mediciones utilizadas en Educación. Se incluyen tres tipos de validez: de *contenido*, *relacionada con el criterio* y de *constructo*. Cada una de éstas se examinará de forma breve, aunque se pondrá un mayor énfasis en la validez de constructo, ya que tal vez sea la forma más importante de validez, desde el punto de vista de la investigación científica.

Validez de contenido y validación de contenido

Una profesora universitaria de psicología ha impartido un curso para estudiantes del último año, donde enfatizó la comprensión de los principios del desarrollo humano. Ella prepara

una prueba de tipo objetivo. Al querer conocer su validez, examina críticamente la relevancia de cada uno de los reactivos de la prueba, para entender los principios del desarrollo humano. Además les pide a dos colegas que evalúen el contenido de la prueba. Naturalmente, les informa a sus colegas lo que está tratando de medir. Ella está investigando la *validez de contenido* de la prueba.

La validez de contenido es la *representatividad* o la *adecuación de muestreo* del contenido —la sustancia, la materia, el tema— de un instrumento de medición. La *validación de contenido* está guiada por la pregunta: ¿la sustancia o contenido de esta medida es representativa del contenido del universo de contenido de la propiedad que se mide? Cualquier propiedad psicológica o educativa posee un universo teórico de contenido, que consiste en todas las posibles cosas que se dicen u observan acerca de la propiedad. Los miembros de este universo, *U*, pueden denominarse "reactivos". La propiedad puede ser el "rendimiento aritmético", por dar un ejemplo relativamente simple. *U* posee un número infinito de miembros: todos los reactivos posibles utilizando números, operaciones aritméticas y conceptos. Una prueba con alta validez de contenido sería teóricamente una muestra representativa de *U*. Si fuera posible elegir aleatoriamente reactivos de *U* en número suficiente, entonces cualquiera de estas muestras de reactivos supuestamente formaría una prueba con una alta validez de contenido. Si *U* comprende los subconjuntos *A*, *B* y *C*, que son operaciones aritméticas, conceptos aritméticos y manipulaciones numéricas, respectivamente, entonces cualquier muestra de *U* lo suficientemente grande representaría a *A*, *B* y *C* de forma casi igual. La validez de contenido de la prueba sería satisfactoria.

Por desgracia, la mayoría de las veces no es posible elegir muestras aleatorias de reactivos de un universo de contenido; dichos universos sólo existen en teoría. Es verdad que es posible y deseable armar grandes grupos de reactivos, especialmente en el área de rendimiento, y obtener muestras aleatorias a partir de dichos grupos, con propósitos de prueba. Pero la validez de contenido de dichos grupos está siempre en duda, no importa qué tan abundantes y qué tan "buenos" sean los reactivos.

Si no es posible satisfacer la definición de validez de contenido, ¿cómo puede lograrse un nivel razonable de validez de contenido? La validación de contenido consiste esencialmente de juicio. Solo o con otros, el investigador juzga la representatividad de los reactivos. Se puede plantear la pregunta: ¿este reactivo mide la propiedad *M*? Expresado de manera más completa, se podría plantear la pregunta: ¿este reactivo es representativo del universo de contenido de *M*? Si *U* tiene subconjuntos, tales como los que se indicaron antes, entonces se deben plantear preguntas adicionales; por ejemplo: ¿este reactivo es miembro del subconjunto M_1 o del subconjunto M_2?

Algunos universos de contenido son más obvios y más fáciles de juzgar que otros; el contenido de muchas pruebas de rendimiento, por ejemplo, parecería obvio. Se dice que puede suponerse la validez de contenido de tales pruebas. Mientras que esta afirmación parece razonable, y mientras el contenido de la mayoría de las pruebas de rendimiento está "autovalidado" en el sentido de que, hasta cierto grado, el individuo que escribe la prueba define la propiedad que se está midiendo (por ejemplo, un maestro que escribe una prueba de ortografía o aritmética para la clase), es peligroso asumir la adecuación de la validez de contenido sin realizar esfuerzos sistemáticos para verificar el supuesto. Por ejemplo, un investigador educativo que comprueba hipótesis acerca de las relaciones entre el rendimiento en estudios sociales y otras variables, puede suponer la validez de contenido de una prueba de estudios sociales. Sin embargo, la teoría a partir de la cual se derivaron las hipótesis quizá requiera *comprensión* y *aplicación* de ideas de estudios sociales; mientras que la prueba utilizada puede tener un contenido casi puramente factual. La prueba carece de validez de contenido en su propósito. De hecho, el investigador no está comprobando en realidad las hipótesis establecidas.

Entonces, la validación de contenido es básicamente de juicio. Los reactivos de una prueba deben estudiarse y se debe ponderar la representatividad supuesta de cada reactivo en el universo, lo cual quiere decir que cada reactivo debe juzgarse respecto a su supuesta relevancia respecto a la propiedad que se mide; no es una tarea fácil. Por lo común, otros jueces "competentes" deben juzgar el contenido de los reactivos. De ser posible, el universo de contenido debe estar claramente definido; es decir, se les deben facilitar a los jueces instrucciones específicas para realizar juicios, así como las especificaciones sobre lo que están juzgando. Después, es posible utilizar algún método para agrupar los juicios independientes. Una excelente guía para la validez de contenido de pruebas de rendimiento es Bloom (1956), quien representa un intento exhaustivo por determinar y discutir objetivos educativos en relación con la medición. El trabajo de Bloom se denominó "taxonomía de Bloom".

Existe otro tipo de validez que es muy similar a la validez de contenido. Ésta se llama *validez aparente* o *de facie*, la cual no es una validez en el sentido técnico; se refiere a aquello que la prueba aparenta medir. Individuos entrenados o sin entrenamiento observarían la prueba y decidirían si ésta mide lo que se supone que debe medir. No se calcula la cuantificación del juicio ni tampoco un índice del acuerdo entre jueces. La validez de contenido es cuantificable a través del empleo de índices de concordancia de las evaluaciones de los jueces. Uno de dichos índices es la *Kappa* de Cohen (Cohen, 1960).

Validez relacionada con el criterio y validación

Como su burdo y desafortunado nombre lo indica, la *validez relacionada con el criterio* se estudia al comparar las puntuaciones de una prueba o escala con una o más variables externas, o criterios, que se sabe o se considera que miden el atributo que se estudia. Un tipo de validez relacionada con el criterio es la llamada *validez predictiva*. El otro tipo es la *validez concurrente*, que difiere de la predictiva en la dimensión del tiempo. La validez predictiva involucra el uso de desempeños del criterio futuros; mientras que la validez concurrente mide el criterio casi al mismo tiempo. En este sentido, la prueba sirve para evaluar el estatus presente del individuo.

La validez concurrente con frecuencia se utiliza para validar una prueba nueva. Para cada examinado se toman por lo menos dos medidas concurrentes. Una de ellas sería la prueba nueva y la otra sería una prueba o medida existente. La validez concurrente se calcularía al correlacionar los dos conjuntos de calificaciones. En el área de las pruebas de inteligencia, las pruebas nuevas e inclusive las revisiones de pruebas antiguas, se utiliza generalmente la prueba de Stanford-Binet o la prueba de Wechsler como criterio concurrente.

Cuando se predice el éxito o fracaso de los estudiantes a partir de sus medidas de aptitud académica, se está considerando la validez predictiva relacionada con el criterio. ¿Qué tan bien predice la prueba (o pruebas) el promedio final o el de la licenciatura? Aquí el enfoque no es tanto lo que la prueba mide, sino su habilidad predictiva. De hecho, en la validación relacionada con el criterio, la cual es con frecuencia investigación práctica y aplicada, el interés básico está más centrado en el criterio, es decir, en los resultados prácticos, que en los predictores. (En la investigación básica esto no es así.) A mayor correlación entre una medida o medidas de aptitud académica y el criterio, por ejemplo la calificación promedio, mejor será la validez. Breve y nuevamente, se enfatiza el criterio y su predicción. Thorndike (1996) ofrece un análisis sobre lo que constituye un buen criterio.

El término *predicción* está generalmente asociado con el futuro. Esto es desafortunado ya que, en la ciencia, predicción no necesariamente significa pronóstico. Se "predice" una

variable dependiente a partir de una variable independiente. Se "predice" la existencia o no-existencia de una relación; ¡incluso se "predice" algo que sucedió en el pasado! Este amplio significado de predicción es el que se utiliza aquí. En cualquier caso, la validez relacionada con el criterio está caracterizada por la predicción sobre un criterio *externo* y por la verificación de un instrumento de medición, ya sea ahora o en el futuro, contra un resultado o medida. En cierto sentido todas las pruebas son predictivas, pues "predicen" cierto tipo de resultado, una situación presente o futura. Las pruebas de aptitud predicen el rendimiento futuro; las pruebas de rendimiento, el rendimiento y competencia presentes y futuras, y las pruebas de inteligencia, la habilidad presente y futura para aprender y resolver problemas. Aun cuando se mide el autoconcepto, se predice que si la puntuación del autoconcepto es tal, entonces el individuo será de tal o cual manera ahora y en el futuro.

La mayor dificultad de la validación relacionada con el criterio es el criterio mismo. Obtener un criterio puede ser incluso difícil. ¿Qué criterio puede utilizarse para validar una medida de eficacia de un profesor? ¿Quién debe juzgar la eficacia de un profesor? ¿Qué criterio puede utilizarse para probar la validez predictiva de una prueba de aptitud musical?

Aspectos de decisión de la validez

Como se indicó antes, la validez relacionada con el criterio está asociada generalmente con resultados y problemas prácticos. El interés no se centra tanto en lo que está detrás del desempeño en la prueba, sino en su utilidad para resolver problemas prácticos y tomar decisiones. Se utilizan cientos de pruebas con los propósitos predictivos de evaluar y seleccionar candidatos potencialmente exitosos en educación, negocios y otras ocupaciones. ¿Ayuda materialmente una prueba o un conjunto de pruebas para decidir sobre la asignación de individuos a empleos, clases, escuelas y otros aspectos similares? Cualquier decisión implica una elección entre tratamientos, asignaciones o programas. Cronbach (1971) señala que para tomar una decisión, se predice el éxito de la persona bajo cada tratamiento y luego se utiliza alguna regla para traducir la predicción en una tarea o recomendación. Una prueba con alta validez relacionada con el criterio ayuda a los investigadores a tomar decisiones exitosas al asignar personas a tratamientos, considerando *tratamientos* en un sentido amplio. Un comité o jefe de admisiones decide si admite o no en la universidad a un solicitante, con base en una prueba de aptitud académica. En efecto, tal uso de las pruebas es bastante importante, y la validez predictiva de las pruebas también tiene gran importancia. Se recomienda el lector al ensayo de Cronbach para una buena exposición de los aspectos de toma de decisión de pruebas y validez.

Taylor y Russell (1939) realizaron una gran contribución en esta área, pues demostraron que las pruebas con poca validez aun pueden utilizarse de manera efectiva con propósito de decisiones. Desarrollaron la tabla Taylor-Russell, que utiliza tres piezas de información: el coeficiente de validez, la tasa de selección y la tasa base. La tasa de selección se refiere al número de personas (solicitantes) que se elegirán del número total de personas. Si hubiera sólo 10 plazas y 100 solicitantes, la tasa de selección sería .10 o 10%. La tasa base es la proporción de personas en la población con ciertas características. Estos datos por lo general se reportan en la prensa. La tasa base de mujeres es, por ejemplo, .52 o 52% de la población de Estados Unidos. Sin utilizar una prueba, si se reúnen aleatoriamente 100 personas en un cuarto, 52 de ellas serían mujeres. Cada uno de los tres componentes puede variar y el hacerlo tiene un efecto sobre la precisión de la selección. Es decir, ayuda a tomar una mejor decisión. Anastasi y Urbina (1997) ofrecen una buena explicación sobre la forma en que funciona este método. El lector interesado necesitará

consultar el artículo original de Taylor y Russell para ver el rango completo de tablas. En esencia, es posible realizar una mejor predicción utilizando una prueba con poca validez si la tasa de selección es pequeña. Desde 1939 este método ha sufrido algunas modificaciones y adiciones, entre las que se incluyen las de Abrahams, Alf y Wolfe (1971); Pritchard y Kazar (1979) y Thomas, Owen y Gunst (1977).

Predictores y criterios múltiples

Se utilizan tanto los predictores múltiples como los criterios múltiples. Más adelante, cuando se estudie la regresión múltiple, se enfocarán los predictores múltiples y la manera de manejarlos estadísticamente. Los criterios múltiples pueden manejarse de forma separada o juntos, aunque esto último no es fácil. En la investigación práctica por lo común debe tomarse una decisión. Si existe más de un criterio, ¿cómo se pueden combinar mejor para tomar una decisión? Por supuesto, debe considerarse la importancia relativa de los criterios. ¿Se desea un administrador con alta habilidad en solución de problemas, con alta habilidad en relaciones públicas o con ambas? ¿Cuál es más importante para un trabajo en particular? Es altamente probable que se haga común el uso tanto de los predictores múltiples como de los criterios múltiples, conforme se comprendan mejor los métodos multivariados y se utilice rutinariamente la computadora en la investigación predictiva.

Validez de constructo y validación de constructo

La validez de constructo es uno de los avances científicos más significativos de la teoría y de la práctica de la medición moderna. Representa un avance significativo ya que liga conceptos y prácticas psicométricos con conceptos teóricos. El trabajo clásico en el área es el de Cronbach y Meehl (1955). Cuando los expertos en medición investigan la validez de constructo de una prueba, casi siempre desean saber qué propiedad o propiedades psicológicas o de otro tipo pueden "explicar" la varianza de las pruebas. Buscan conocer el "significado" de las pruebas. Si se trata de una prueba de inteligencia, ellos desean saber qué factores subyacen al desempeño en la prueba. Plantean la pregunta: ¿qué factores o constructos explican la varianza del desempeño en la prueba? ¿Esta prueba mide habilidad verbal y habilidad de razonamiento abstracto? ¿"Mide" también la pertenencia a una clase social? Ellos preguntan, por ejemplo, qué proporción de la varianza total de la prueba es explicada por cada uno de los constructos como habilidad verbal, habilidad de razonamiento abstracto y pertenencia a una clase social. En síntesis, buscan explicar las diferencias individuales en las puntuaciones de la prueba. Su interés por lo general está centrado en las propiedades que se miden, más que en las pruebas utilizadas para lograr la medición.

Los investigadores por lo común inician con los constructos o variables que tienen relación. Suponga que un investigador ha descubierto una correlación positiva entre dos medidas: una de tradicionalismo educativo y la otra sobre la percepción de las características asociadas con un "buen" profesor. Los individuos con puntuaciones altas en la medida de tradicionalismo ven al "buen" profesor como eficiente, moral, minucioso, industrioso, concienzudo y confiable. Los individuos con puntuaciones bajas en la medida de tradicionalismo quizá vean al "buen" profesor de una forma diferente. El investigador ahora desea saber *por qué* existe dicha relación, es decir, lo que está detrás de ella. Para lograr esto, debe estudiarse el significado de los constructos incluidos en la relación: "percepción del 'buen' maestro" y "tradicionalismo". La manera de estudiar estos significados implica un problema de validez de constructo. Este ejemplo fue tomado de Kerlinger y Pedhazur (1968).

Se puede ver que la validación de constructo y la investigación científica empírica están íntimamente relacionadas. No es simplemente cuestión de validación de una prueba.

Debe intentarse validar la teoría que está detrás de la prueba. Cronbach (1990) indica que existen tres partes en la validación de constructo: sugerir qué constructos posiblemente explican el desempeño en la prueba, derivar hipótesis a partir de la teoría que incluye al constructo y comprobar empíricamente las hipótesis. Tal planteamiento es una precisión del modelo científico general analizado en capítulos anteriores.

El aspecto más importante sobre la validez de constructo que la separan de otros tipos de validez es su preocupación por la teoría, los constructos teóricos y la investigación científica empírica, incluyendo la comprobación de relaciones hipotetizadas. La validación de constructo en medición contrasta en forma notable con modelos que definen la validez de una medida, principalmente por su éxito al predecir el criterio. Por ejemplo, un aplicador de pruebas puramente empírico podría decir que una prueba es válida si distingue de manera eficiente entre individuos con altos o bajos niveles de cierto rasgo. El *porqué* de que la prueba sea exitosa al separar los subconjuntos de un grupo no tiene gran importancia. Es suficiente con que lo haga.

Convergencia y discriminación

Observe que la comprobación de hipótesis alternativas es particularmente importante en la validación de constructo, ya que se requiere tanto de la convergencia como de la discriminación. *Convergencia* significa que la evidencia de diferentes fuentes, reunida de diferentes maneras, indica un significado similar o igual al del constructo. Diferentes métodos de medición deben convergir en el constructo. La evidencia producida al aplicar el instrumento de medición a diferentes grupos en diferentes lugares debe producir significados similares o, si no es así, entonces debe explicar las diferencias. Por ejemplo, una medida del autoconcepto de niños debe ser capaz de ofrecer interpretaciones similares en distintas partes del país. Si no es capaz de ofrecer dichas interpretaciones en cierta localidad, entonces la teoría debe ser capaz de explicar por qué —de hecho debe predecir tal diferencia—.

Discriminación significa que es posible diferenciar empíricamente el constructo de otros constructos que puedan ser similares, y que se puede señalar lo que ***no está relacionado*** con el constructo. En otras palabras, se señala qué otras variables están correlacionadas con el constructo y de qué manera lo están. Sin embargo, también se indica cuáles variables no deben estar correlacionadas con el constructo. Por ejemplo, se señala que una escala para medir el *conservadurismo* debe correlacionarse sustancialmente, y de hecho lo hace, con medidas de *autoritarismo* y *rigidez* —la teoría predice esto— pero no se correlaciona con medidas de *aceptación social* (véase Kerlinger, 1970). A continuación se ejemplificarán estas ideas.

Un ejemplo hipotético de validación de constructo

Suponga que un investigador está interesado en los determinantes de la creatividad y la relación de la creatividad con el rendimiento escolar. El investigador nota que las personas más sociables, quienes muestran afecto hacia otros, también parecen ser menos creativos que aquellos que son menos sociables y afectuosos. El objetivo consiste en probar la relación implicada de una manera controlada. Una de las primeras tareas es obtener o construir una medida de la característica sociable-afectuoso. El investigador, conjeturando que esta combinación de rasgos quizá sea un reflejo de un interés más profundo en el amor por los demás, lo llama *amorismo*. Se asume que existen diferencias individuales respecto al amorismo, es decir, algunas personas lo poseen en gran cantidad, otras en cantidad moderada y otras muy poco.

El primer paso es construir un instrumento para medir el amorismo. La literatura ofrece poca ayuda, puesto que los psicólogos científicos han estudiado muy poco la naturaleza fundamental del amor. No obstante, se ha medido la sociabilidad. El investigador debe construir un nuevo instrumento, basando su contenido en conceptos intuitivos y racionales sobre lo que es el amorismo. La confiabilidad de la prueba, que fue probada con grupos grandes, oscila entre .75 y .85.

La pregunta ahora es si la prueba es o no válida. El investigador correlaciona el instrumento y lo llama escala *A*, con las medidas independientes de sociabilidad. Las correlaciones son moderadamente altas, pero se necesita mayor evidencia para afirmar que la prueba posee validez de constructo. Se deducen ciertas relaciones que deben existir o no entre el amorismo y otras variables. Si el amorismo es la tendencia general de amar a los demás, entonces debe correlacionarse con características tales como ser cooperativo y amistoso. Las personas con alto amorismo enfrentarán los problemas de una forma orientada al *yo*; en contraste con las personas con bajo amorismo, quienes enfrentarán los problemas de una forma orientada a la tarea.

Con base en este razonamiento, el investigador aplica la escala *A* y una escala para medir subjetividad a un grupo de estudiantes del primer año de preparatoria. Para medir el nivel de cooperación se realiza una observación del comportamiento del mismo grupo de estudiantes en el salón de clase. Las correlaciones entre las tres medidas son positivas y altas. Observe que no se esperaría una correlación alta entre las medidas. Si las correlaciones fueran demasiado altas, entonces se dudaría con respecto a la validez de la escala *A*; quizás estaría midiendo subjetividad o nivel de cooperación, pero no amorismo.

Debido a que conoce las desventajas de la medición psicológica, el investigador no está satisfecho. Estas correlaciones positivas tal vez se deban a un factor común a las tres pruebas, pero irrelevante para el amorismo; por ejemplo, la tendencia a dar respuestas "correctas" o deseables. (Sin embargo, esto podría descartarse a causa de que la medida de observación del cooperativismo se correlaciona positivamente con el amorismo y la subjetividad.) Por lo tanto, con un nuevo grupo de participantes, el investigador aplica las escalas de amorismo y subjetividad, evalúa la conducta de cooperativismo de los participantes y, además, aplica una prueba de creatividad que demostró ser confiable en otra investigación.

El investigador establece la relación entre amorismo y creatividad en la forma de una hipótesis: la relación entre la escala *A* y la medida de creatividad será negativa y significativa. Las correlaciones entre amorismo y cooperativismo, y entre amorismo y subjetividad serán positivas y significativas. También se formulan hipótesis de "verificación": la correlación entre cooperativismo y creatividad no será significativa, será cercana a cero; pero la correlación entre subjetividad y creatividad será positiva y significativa. Esta última relación se predice con base en hallazgos previos de investigación. Los seis coeficientes de correlación se presentan en la matriz de correlación de la tabla 28.1. Las cuatro medidas se denominan de la siguiente forma: A, amorismo; B, cooperativismo; C, subjetividad, y D, creatividad.

La evidencia de la validez de constructo de la escala *A* es buena. Todas las *r* resultaron tal como se predijo; de especial importancia son las *r* entre D (creatividad) y las otras variables. Note que hay tres tipos diferentes de predicción: positiva, negativa y cero; las tres resultaron tal como se predijo. Lo anterior ilustra lo que se llamaría *predicción diferencial* o *validez diferencial* —o discriminación—. No es suficiente predecir, por ejemplo, que la medida que se supone refleja la propiedad estudiada esté correlacionada en forma positiva con una variable teóricamente relevante. Se debería, deduciendo a partir de la teoría, predecir más de una de dichas relaciones positivas. Además, deberían predecirse relaciones de cero entre la variable principal y las variables "irrelevantes" con la teoría. En el

TABLA 28.1 *Intercorrelaciones de cuatro medidas hipotéticas (N = 90)*[a]

	B	C	D
A	.50	.60	−.30
B		.40	.05
C			.50

[a] A = Amorismo; B = Cooperativismo; C = Subjetividad; D = Creatividad. Los coeficientes de correlación de .25 o mayores son significativos al nivel .01.

ejemplo anterior, aunque se esperaba que el cooperativismo se correlacionara con el amorismo, no hubo una razón teórica para esperar que se correlacionara en lo absoluto con la creatividad.

Un ejemplo de diferente tipo es el investigador que introduce deliberadamente una medida que invalidaría otras relaciones positivas, si dicha variable se correlaciona con la variable cuya validez se estudia. Un gran problema de las escalas de personalidad y de actitud es el fenómeno que involucra el deseo de ser aceptado socialmente, que se mencionó antes. La correlación entre la variable estudiada y una variable teóricamente relacionada tal vez se deba a que ambos instrumentos estén midiendo el deseo de aceptación social más que las variables para las que fueron diseñados. Dicha tendencia se verifica, en parte, si se incluye una medida del deseo de aceptación social junto con otras medidas.

A pesar de que todas las evidencias conduzcan al investigador a creer que la escala *A* posee validez de constructo, aún pueden existir dudas. Por lo tanto, se realiza un estudio donde los alumnos con alto y bajo nivel de amorismo deben resolver problemas. Se predice que los alumnos con bajo nivel de amorismo resolverán los problemas con más éxito que aquellos con alto amorismo. Si los datos apoyan la predicción, esto representa mayor evidencia de la validez de constructo de la medida de amorismo. Esto es, por supuesto, un hallazgo significativo en sí mismo. No obstante, probablemente un procedimiento como éste sea más apropiado para medidas de rendimiento y de actitud. Por ejemplo, es posible manipular las comunicaciones para cambiar actitudes. Si las puntuaciones de actitud cambian de acuerdo con la predicción teórica, entonces ello sería evidencia de la validez de constructo de la medida de actitud, ya que las puntuaciones quizá no cambiarían de acuerdo con la predicción si la medida no estuviera midiendo el constructo.

El método multirrasgo–multimétodo

Una contribución significativa e influyente de Campbell y Fiske (1959) en la comprobación de la validez es el empleo de las ideas de convergencia y discriminación y de matrices de correlación, para aportar evidencia sobre la validez. Para explicar el método se usarán algunos datos de un estudio sobre actitudes sociales de Kerlinger (1967, 1984). Se ha encontrado que existen dos dimensiones básicas de las actitudes sociales, que corresponden a descripciones filosóficas, sociológicas y políticas del liberalismo y conservadurismo. Se aplicaron dos tipos de escalas diferentes a estudiantes de educación de posgrado y a grupos fuera de las universidades en Nueva York, Texas y Carolina del Norte. Un instrumento, la Escala de Actitudes Sociales, contenía *afirmaciones* usuales de actitud: 13 reactivos liberales y 13 conservadores. El segundo instrumento, Referentes-I o REF-I, utilizaba *referentes* de actitud (palabras y frases cortas: *propiedad privada*, *religión* y *derechos civiles*, por ejemplo) como reactivos, de los cuales 25 eran referentes liberales y 25 eran referentes conservadores.

Las muestras, las escalas y parte de los resultados se describen en Kerlinger (1972). Los datos reportados en la tabla 28.2 fueron obtenidos de una muestra de Texas, $N = 227$ estudiantes de posgrado.

Entonces, se tienen dos tipos de instrumentos de actitud completamente diferentes: uno con reactivos de referencia y el otro con reactivos afirmativos, o método 1 y método 2, respectivamente. Las dos dimensiones básicas medidas fueron el liberalismo (L) y el conservadurismo (C). ¿Miden liberalismo y conservadurismo las subescalas L y C de las dos escalas? Parte de la evidencia se muestra en la tabla 28.2, la cual presenta la correlación entre las cuatro subescalas de los dos instrumentos, así como los coeficientes de confiabilidad de la subescala, calculados a partir de las respuestas a las dos escalas.

En un análisis multirrasgo-multimétodo se utiliza más de un atributo y más de un método en el proceso de validación. Los resultados de correlacionar variables dentro y entre métodos pueden presentarse en la llamada matriz multirrasgo-multimétodo. La matriz presentada en la tabla 28.2 es la forma más simple posible de realizar un análisis de este tipo: dos variables y dos métodos. Por lo común se desearía utilizar más variables.

La parte más importante de la matriz es la diagonal que contiene las correlaciones entre los métodos; en la tabla 28.2 este resultado se ubica en la sección inferior izquierda de la tabla. Los valores diagonales deben ser sustanciales, pues reflejan las magnitudes de las correlaciones entre las mismas variables, medidas de forma distinta. Estos valores, expresados en itálicas en la tabla (.53 y .54) son bastante altos.

En este ejemplo, la teoría exige correlaciones cercanas a cero o correlaciones bajas negativas entre L y C (véase Kerlinger, 1967 para mayor profundidad sobre esto). La correlación entre L_1 y C_1 es –.07 y entre L_2 y C_2 es –.09, lo cual coincide con la teoría. La correlación cruzada entre L y C, es decir, la correlación entre L del método 1 y C del método 2, o entre L_1 y C_2, es –.37, mayor de lo que la teoría predice (se adoptó un límite superior de –.30). Entonces, con excepción de la correlación cruzada de –.37 entre L_1 y C_2, se sostiene la validez de constructo de la escala de actitudes sociales. Por supuesto que se desearía mayor evidencia que los resultados obtenidos con una muestra, y que también se desearía una explicación respecto a la alta correlación negativa de método cruzado entre L_1 y C_2. No obstante, el ejemplo ilustra las ideas básicas del método múltirrasgo-multimétodo para la validez.

Campbell y Fiske (1959) utilizaron terminología específica para describir cada correlación en la tabla. Las correlaciones *monométodo-monorrasgo* son las confiabilidades. Éstas se encuentran en la diagonal principal de la matriz; en la tabla 28.2 son los valores .85, .88, .81 y .82 encerrados en paréntesis. Las correlaciones *heterométodo-monorrasgo* representan

TABLA 28.2 *Correlaciones entre dimensiones de actitudes sociales a través de dos métodos de medición, modelo multirrasgo-multimétodo, muestra de Texas ($N = 227$)*[a]

		Método 1 (Referentes)		*Método 2 (Afirmaciones)*	
		L_1	C_1	L_2	C_2
Método 1	L_1	(.85)			
(Referentes)	C_1	–.07	(.88)		
Método 2	L_2	*.53*	–.15	(.81)	
(Afirmaciones)	C_2	–.37	*.54*	–.09	(.82)

[a] Método 1: referentes; método 2: afirmaciones; L = liberalismo; C = conservadurismo. Las cifras en paréntesis sobre la diagonal son índices de confiabilidad de la consistencia interna; las cifras en itálicas (.53 y .54) son correlaciones del cruce de los métodos L-L y C-C (validez).

la validez que se analizó anteriormente, que son los valores .53 y .54 escritos en itálicas en la tabla 28.2. Existen otros dos tipos de correlación: la *monométodo-heterorrasgo* (los valores –.07 y –.09), y la *heterométodo-heterorrasgo* (que fueron –.37 y –.15). Campbell y Fiske afirman que para obtener evidencia completa de la validez de constructo, las correlaciones deben seguir un patrón establecido. Si no se logran cubrir los requisitos se debilitan los aspectos de la validez. Algunos artículos han intentado resolver este problema al relajar algunos de los requisitos. Tales artículos afirman haber logrado un grado de éxito parcial.

El modelo del método multirrasgo-multimétodo constituye un ideal. Si es posible debe realizarse. En realidad la investigación y la medición de constructos importantes como el conservadurismo, la agresividad, la calidez del profesor, la necesidad de rendimiento, la honestidad, etcétera, finalmente lo requieren. Sin embargo, en muchas situaciones de investigación es difícil o aun imposible aplicar dos o más medidas de dos o más variables con muestras relativamente grandes. Aunque siempre deben realizarse esfuerzos para estudiar la validez, la investigación no debe abandonarse sólo porque no es posible aplicar el método completo.

Ejemplos de investigación de la validación concurrente

Wood (1994) ofrece un buen ejemplo de cómo validar una prueba que utiliza datos médicos y psicológicos. Aquí el criterio es una medición física real. Wood desarrolló un instrumento llamado instrumento de evaluación de la eficiencia del autoexamen de mama (Breast Self-Examination Proficiency Rating Instrument, BSEPRI), el cual mide qué tanto conocimiento tiene quien toma la prueba, respecto al autoexamen de mama. Las participantes en el estudio eran estudiantes de enfermería. A la mitad de ellas se les dieron instrucciones sobre el autoexamen y a la otra mitad no. Una prueba *t* demostró que quienes recibieron instrucciones obtuvieron puntuaciones significativamente mayores que quienes no las recibieron. Wood obtuvo la validez concurrente al correlacionar las puntuaciones de palpación del instrumento con la habilidad de los estudiantes para detectar protuberancias en un modelo de silicón.

Iverson, Guirguis y Green (1998) examinaron la validez concurrente en una forma breve de la escala Wechsler de inteligencia para adultos-revisada (WAIS-R). Esta forma breve consistía de siete escalas. Fue desarrollada para evaluar pacientes con diagnóstico de un trastorno del espectro de la esquizofrenia. Las puntuaciones del CI calculadas por medio de esta forma breve tienen una alta correlación con las puntuaciones del CI de la escala completa. Los CI verbales, los CI de ejecución y los CI de la escala completa, calculados con la forma breve, estaban altamente correlacionados con los CI de la escala completa. Las correlaciones (coeficientes de validez) oscilaron entre .95 y .98. En general, la forma breve de siete subescalas mostró validez concurrente adecuada y sirve para evaluar el funcionamiento intelectual de personas con trastornos psicóticos. Iverson y colaboradores correlacionaron la prueba nueva (forma breve) con la prueba establecida (escala completa) para obtener una medida de validez concurrente. Comrey (1993) utilizó un procedimiento similar para crear la forma breve de las escalas de personalidad de Comrey (Comrey Personality Scales, CPS). Con el uso de datos ya existentes Comrey extrajo los "mejores" reactivos de cada escala (que se analizan más adelante) y calculó dos puntuaciones totales: una de la forma breve y otra de la forma original. La correlación de las dos puntuaciones produjo un valor de validez concurrente.

Ejemplos de investigación de validación de constructo

En cierto sentido, cualquier tipo de validación es validación de constructo. Loevinger (1957) argumenta que la validez de constructo, desde un punto de vista científico, constituye

el total de la validez. En el otro extremo, Bechtoldt (1959) argumenta que la validez de constructo no tiene lugar en la psicología. Horst (1966) dice que es muy difícil aplicar las ideas de Cronbach y Meehl dentro de la teoría lógica y práctica de la psicometría. Sin embargo, cuando se prueban hipótesis y cuando se estudian relaciones empíricamente, se involucra la validez de constructo. Debido a su importancia, ahora se examinarán dos ejemplos de investigación sobre la validación de constructo.

Una medida de antisemitismo

En un intento inusual por validar su medida sobre antisemitismo, Glock y Stark (1966) utilizaron las respuestas a dos frases incompletas respecto a los judíos: "Es una pena que los judíos..." y "No puedo entender por qué los judíos..." Quienes calificaron las frases consideraron lo que cada sujeto había escrito y caracterizaron las respuestas como imágenes negativas, neutrales o positivas sobre los judíos. Entonces, cada sujeto fue considerado individualmente como poseedor de una de tres imágenes diferentes sobre los judíos. Cuando las respuestas al índice de creencias antisemitas (Index of Anti-Semitic Beliefs), la medida que se estaba validando, se dividieron en sin-antisemitismo, antisemitismo medio, antisemitismo medio alto y antisemitismo alto, los porcentajes de respuestas negativas a las dos frases incompletas fueron, respectivamente: 28, 41, 61, 75. Esto representa una buena evidencia de validez, ya que los individuos categorizados desde sin-antisemitismo hasta antisemitismo alto por medio de la medida a ser validada, el índice de creencias antisemitas, respondieron a una medida completamente diferente de antisemitismo, las dos con frases incompletas, de manera congruente con su categorización por medio del índice.

Una medida de personalidad

En un capítulo posterior se discutirá una importante herramienta analítica llamada *análisis factorial*. No obstante, es necesario mencionar este método para la comprensión de la validación de constructo. En años recientes, el análisis factorial parece ser el método de elección para muchas personas involucradas con la validez de constructo. El análisis factorial es esencialmente un método para encontrar aquellas variables que tienen algo en común. Si algunos reactivos de una prueba de personalidad están diseñados para medir extroversión, entonces, en un análisis factorial, dichos reactivos deben cargarse mucho hacia un factor y poco hacia los otros.

A mediados de los años cincuenta, el profesor Andrew L. Comrey, de la Universidad de California en Los Ángeles, realizó la tarea de examinar todas las pruebas de personalidad publicadas reconocidas. Su objetivo inicial era tratar de determinar cuál era la medida correcta (válida) de personalidad. Para esto, el doctor Comrey utilizó un análisis factorial. Contrariamente a sus expectativas iniciales, surgió una nueva prueba de personalidad de carácter único. La prueba de personalidad de Comrey, ahora conocida como las escalas de personalidad de Comrey (Comrey Personality Scales) (CPS), fue de las primeras pruebas desarrolladas por medio del uso del análisis factorial. En 1970, después de un proceso de 15 años de investigación y construcción de la prueba, se publicaron las escalas de personalidad de Comrey (véase Comrey y Lee, 1992 para encontrar un resumen y el procedimiento). El constructo de Comrey de personalidad consta de ocho dimensiones principales:

Confianza contra defensividad
Disciplina contra falta de compulsión
Conformismo social contra rebeldía
Actividad contra falta de energía

Estabilidad emocional contra neuroticismo
Extroversión contra introversión
Masculinidad contra feminidad (renombrados dureza mental contra sensibilidad)
Empatía contra egocentrismo

Desde 1970, Comrey ha publicado diversos artículos que apoyan la validez de sus escalas de personalidad. Esto se hizo al aplicar las CPS, o una forma traducida de las CPS, a diferentes grupos de personas. Después de obtener los datos, cada grupo de éstos fue analizado factorialmente. En cada caso surgieron los mismos ocho factores. Aunque esto no afirma que existan exclusivamente ocho factores de personalidad, los datos lo sustentan. En una investigación reciente realizada por Brief, Comrey y Collins (1994), las CPS fueron traducidas al ruso y aplicadas a 287 participantes hombres y 170 participantes mujeres. Los datos apoyaron seis de las ocho subescalas. Las únicas subescalas que no recibieron suficiente apoyo fueron la de Empatía contra Egocentrismo y la de Actividad contra Falta de Energía.

En un artículo breve, Comrey, Wong y Backer (1978) presentan un procedimiento simple para validar la escala de Conformidad Social contra Rebeldía. En un estudio, Comrey y colaboradores reclutaron a dos grupos de participantes: asiáticos y no-asiáticos. La percepción tradicional de los asiáticos es que son más conformistas socialmente que los no-asiáticos. Existe alguna evidencia que apoya esta afirmación, tal como una fuerte influencia paterna, fuertes valores tradicionales, etcétera. [El estudio de Scattone y Saetermoe (1997) es uno de los que ha demostrado lo anterior.] Por lo tanto, en el estudio de Comrey y colaboradores, la idea establecida respecto a la diferencia entre asiáticos y no-asiáticos sobre conformismo social fue utilizada como el criterio o "medida externa". Todos los participantes respondieron las escalas de personalidad de Comrey, aunque sólo la subescala de Conformismo Social contra Rebeldía era de interés para dicho estudio. Con el uso de una prueba *t*, estos investigadores demostraron una diferencia estadísticamente significativa entre asiáticos y no-asiáticos en la escala de Conformismo Social contra Rebeldía. El estudio podría utilizarse como ejemplo para ilustrar la validez discriminante.

El segundo estudio de este artículo demostró la validez convergente. Se espera que la Conformidad Social esté relacionada con la afiliación y filosofía políticas. Generalmente se piensa que los conservadores son más conformistas socialmente que los liberales, a quienes se considera más rebeldes. En este estudio algunas personas completaron las escalas de personalidad de Comrey y respondieron preguntas respecto a su afiliación política. Comrey y colaboradores encontraron una correlación estadísticamente significativa entre la afiliación política y las puntuaciones en la escala de Conformismo Social contra Rebeldía, lo cual proporcionó información adicional respecto a la validez de esa escala. A pesar de que este artículo es breve, está bien presentado. El estudiante aprenderá mucho con la lectura del artículo.

Medición de la democracia

¿Qué quiere decir *democracia*? El término se utiliza constantemente. ¿Pero qué se quiere decir cuando se usa? Aún más difícil, ¿cómo se mide? Bollen (1980) definió y midió "democracia", la utilizó como variable y demostró la validez de constructo de su índice de democracia política (Index of Political Democracy). Él examinó con sumo cuidado sus usos y definiciones previas, explicó la teoría subyacente al constructo y extrajo de medidas anteriores facetas importantes de la democracia política para construir su medida. Ésta contiene dos grandes aspectos —libertad política y soberanía popular— los cuales pueden llamarse variables latentes. Cada aspecto tiene tres facetas: *libertad de prensa*, *libertad de oposición de grupo* y *sanción gubernamental* (ausencia de) por libertades políticas; y *elecciones*

justas, selección ejecutiva y selección legislativa para la soberanía popular. Estos seis "indicadores" se utilizan para medir la democracia política de los países. Cada indicador está definido operacionalmente y se utiliza una escala de 4 puntos para aplicarlos a cualquier nación. La soberanía popular, por ejemplo, se mide al evaluar en qué grado la élite de un país representa al pueblo: derecho del voto, igual peso de los votos y proceso electoral justo. Los seis indicadores se combinan en un índice o puntuación única (véase Bollen, 1979, para una descripción detallada del índice y su puntuación). Note que "indicador" o "indicador social" es un término importante en la investigación social contemporánea. Por desgracia existe poco acuerdo respecto a cuáles son exactamente los indicadores. Se han definido de varias formas como índices de condiciones sociales, estadísticos e incluso como variables. En el artículo de Bollen se consideran variables. Para un análisis sobre las definiciones véase a Jaeger (1978).

A través del análisis factorial y otros procedimientos, Bollen encontró evidencia empírica para apoyar la confiabilidad y la validez de constructo del índice. Él demostró, por ejemplo, que los seis indicadores son manifestaciones de una variable latente subyacente, que es la "democracia política". También demostró que el índice está altamente correlacionado con otras medidas de democracia. Finalmente, se calcularon valores del índice para un gran número de países. Estos valores parecen coincidir con el grado de democracia (en una escala de 0 a 100) de los países; por ejemplo, Estados Unidos, 92.4; Canadá, 99.5; Cuba, 5.2; República de Estados Árabes, 38.7; Suecia, 99.9; Unión Soviética, 18.2; Israel, 96.8. Evidentemente Bollen logró medir con éxito un constructo en extremo complejo y difícil.

Otros métodos de validación de constructo

Además del método multirrasgo-multimétodo y de los métodos utilizados en los estudios anteriores, existen otros métodos para la validación de constructo. Cualquiera que aplique pruebas está familiarizado con la técnica de correlación de los reactivos con las puntuaciones totales. Al usar la técnica se supone que la puntuación total es válida. El reactivo es válido de acuerdo con el grado en que éste mida lo mismo que la puntuación total (véase capítulo 27 o Friedenberg para el estudio del análisis de reactivos).

Para estudiar la validez de constructo de cualquier medida, siempre es útil correlacionar la medida con otras medidas. El ejemplo sobre el amorismo analizado antes ilustró el método y las ideas que están detrás de él. Sin embargo, ¿no sería más valioso correlacionar una medida con un gran número de otras medidas? ¿Existe una mejor manera de aprender sobre un constructo que conocer sus correlatos? El análisis factorial constituye un método refinado para hacer esto, pues indica, en efecto, qué medidas miden la misma cosa y en qué grado miden aquello que miden.

El análisis factorial es un método poderoso e indispensable de la validación de constructo. Bollen (1980) lo utilizó en la validación del índice de democracia política y Comrey lo empleó para desarrollar una prueba completa de personalidad. Aunque ya fue descrito brevemente antes y se estudiará en detalle en un capítulo posterior, su gran importancia para la validación de medidas hace obligatorio describirlo aquí. Se trata de un método para reducir un gran número de medidas a un número más pequeño, llamadas *factores*, al descubrir cuáles "van juntas" (por ejemplo, cuáles miden la misma cosa) y las relaciones entre los grupos de medidas que van juntas. Por ejemplo, se pueden aplicar 20 pruebas a un grupo de individuos, suponiendo que cada una mide algo diferente. Sin embargo, quizá se encuentre que estas 20 pruebas son lo suficientemente redundantes como para ser explicadas con sólo cinco medidas o factores.

Una definición de validez en términos de varianza: la relación de la varianza entre la confiabilidad y la validez

El tratamiento de varianza de la validez presentado aquí es una extensión del tratamiento de confiabilidad presentado en el capítulo 27. Ambos tratamientos siguen la presentación de Guilford de la validez.

En el capítulo anterior la confiabilidad se definió como

$$r_{tt} = \frac{V_{\infty}}{V_t} \tag{28.1}$$

que es la proporción de la varianza "verdadera" entre la varianza total. Es teórica y empíricamente útil definir la validez de forma similar:

$$Val = \frac{V_{\infty}}{V_t} \tag{28.2}$$

donde *Val* es la validez, V_{∞} la varianza del factor común y V_t la varianza total de la medida. Por lo tanto, la validez se considera como la proporción de la varianza total de una medida, que es varianza del factor común.

Por desgracia, todavía no es posible presentar el significado completo de dicha definición, ya que se requiere de la comprensión de la llamada teoría factorial y ésta no se estudiará sino hasta después en el presente libro. A pesar de esta dificultad debe intentarse una explicación de la validez en términos de varianza para lograr una visión completa del tema. Además, la expresión matemática de la validez y la confiabilidad unificará y aclarará ambos temas. De hecho, la confiabilidad y la validez se considerarán como partes de un todo unificado.

La varianza del factor común es la varianza de una medida que es compartida por otras medidas. En otras palabras, la varianza del factor común es la varianza que dos o más pruebas tienen en común.

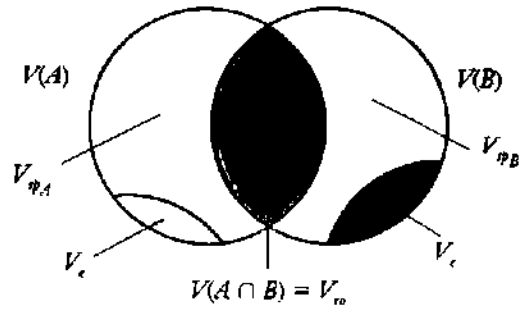
En contraste con la varianza del factor común de una medida está su *varianza específica*, V_{σ}, la varianza sistemática de una medida que no es compartida por cualquier otra medida. Si una prueba mide habilidades que miden otras pruebas, entonces se tiene varianza de factor común; si también mide habilidades que ninguna otra prueba mide, entonces se tiene varianza específica. La figura 28.1 expresa tales ideas y también añade el concepto de la varianza del error. Los círculos A y B representan las varianzas de las pruebas A y B. La intersección de A y B, $A \cap B$, es la relación de los dos conjuntos. De forma similar, $V(A \cap B)$ es la varianza del factor común. También se indican las varianzas específicas y las varianzas del error de ambas pruebas.

Entonces, desde este punto de vista y siguiendo el razonamiento sobre la varianza bosquejado en el capítulo anterior, cualquier varianza total de una medida posee varios componentes: *varianza del factor común*, *varianza específica* y *varianza del error*, lo cual se expresa en la ecuación:

$$V_t = V_{\infty} + V_{\sigma} + V_e \tag{28.3}$$

Para tener la capacidad de hablar de proporciones de la varianza total, se dividen los términos de la ecuación 28.3 entre la varianza total:

FIGURA 28.1



$$\frac{V_t}{V_t} = \frac{V_{vc}}{V_t} + \frac{V_{ve}}{V_t} + \frac{V_e}{V_t} \tag{28.4}$$

¿Cómo es que las ecuaciones 28.1 y 28.2 embonan aquí? El primer término a la derecha del signo de igual, V_{vc}/V_t es el miembro derecho de la ecuación (28.2). Por lo tanto, la validez puede ser considerada como esa parte de la varianza total de una medida que no es varianza específica ni varianza del error, lo cual en forma algebraica se observa así:

$$\frac{V_{vc}}{V_t} = \frac{V_t}{V_t} - \frac{V_{ve}}{V_t} - \frac{V_e}{V_t} \tag{28.5}$$

Por medio de la definición dada en el capítulo anterior, la confiabilidad puede definirse como:

$$r_{tt} = 1 - \frac{V_e}{V_t} \tag{28.6}$$

Lo que puede escribirse como:

$$r_{tt} = \frac{V_t}{V_t} - \frac{V_e}{V_t} \tag{28.7}$$

Sin embargo, la parte derecha de la ecuación es parte del término de la derecha de la ecuación (28.5). Si se modifica la ecuación (28.5) ligeramente, se obtiene:

$$\frac{V_{vc}}{V_t} = \frac{V_t}{V_t} - \frac{V_e}{V_t} - \frac{V_{vp}}{V_t} \tag{28.8}$$

Esto debe significar, entonces, que la validez y la confiabilidad son relaciones de varianza cercanas. La confiabilidad es igual a los primeros dos miembros de la derecha de (28.8). Por lo tanto, al incorporar (28.1) resulta:

$$r_{tt} = \frac{V_t}{V_t} - \frac{V_e}{V_t} = \frac{V_\infty}{V_t} \tag{28.9}$$

Si se sustituye en (28.8), se obtiene:

$$\frac{V_\omega}{V_t} = \frac{V_\infty}{V_t} - \frac{V_\sigma}{V_t} \tag{28.10}$$

De esta forma se observa que la proporción de la varianza total de una medida es igual a la proporción de la varianza total que es varianza "verdadera", menos la proporción que es varianza específica. O bien, la validez de una medida es esa porción de la varianza total de la medida, que comparte varianza con otras medidas. Teóricamente la varianza válida no incluye varianza debida al error, ni tampoco incluye varianza que sea específica únicamente a esta medida.

Todo esto puede resumirse de dos maneras. Primero, se suma en una ecuación o dos. Suponga que se tiene un método para determinar la varianza (o varianzas) del factor común de una prueba. (Posteriormente se verá que el análisis factorial es dicho método.) Para simplificar, considere que hay dos fuentes de varianza del factor común en una prueba —y ninguna otra—. Llame a estos factores *A* y *B*, que pueden ser habilidad verbal y habilidad aritmética, o tal vez actitudes liberales y actitudes conservadoras. Si se añade la varianza de *A* a la varianza de *B*, se obtiene la varianza del factor común de la prueba, la cual se expresa por medio de las ecuaciones:

$$V_\omega = V_A + V_B \tag{28.11}$$

$$\frac{V_\omega}{V_t} = \frac{V_A}{V_t} + \frac{V_B}{V_t} \tag{28.12}$$

Entonces, utilizando (28.2) y sustituyendo en (28.12), se obtiene:

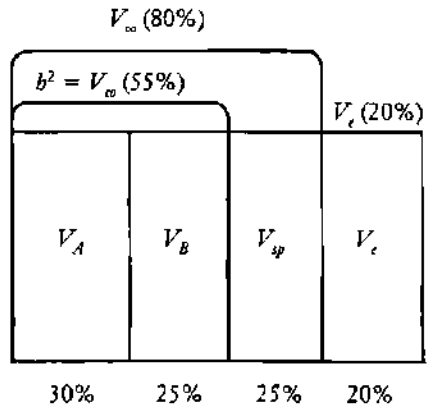
$$Val = \frac{V_A}{V_t} + \frac{V_B}{V_t} \tag{28.13}$$

La varianza total de una prueba, como se indicó antes, incluye la varianza del factor común, la varianza específica para la prueba y no para otra prueba (por lo menos en lo que se refiere a la presente información) y la varianza del error. Las ecuaciones 28.3 y 28.4 así lo expresan. Al sustituir en (28.4) la igualdad de (28.12) se obtiene:

$$\frac{V_t}{V_t} = \underbrace{\overbrace{\frac{V_A}{V_t} + \frac{V_B}{V_t}}^{h_2} + \frac{V_{es}}{V_t}}_{r_{tt}} + \frac{V_e}{V_t} \tag{28.14}$$

Los primeros dos términos del lado derecho de (28.14) están asociados con la validez de la medida, y los primeros tres términos de la derecha están asociados con la confiabilidad de la medida. Estas relaciones ya se han indicado. La varianza del factor común o el componente de validez de la medida se denomina h^2 (*aspectos comunes*), un símbolo que por lo común se utiliza para indicar la varianza del factor común de una prueba. Como siempre, la confiabilidad se denomina r_{tt}.

FIGURA 28.2



Comentar todas las implicaciones de esta formulación de validez y confiabilidad desviaría demasiado el tema en este momento. Todo lo que se necesita ahora es intentar aclarar la formulación con un diagrama y un breve análisis.

La figura 28.2 representa un intento por expresar la ecuación 28.14 en forma de diagrama. La figura indica la contribución de las distintas varianzas a la varianza total (considerada igual al 100%). Cuatro varianzas, tres varianzas sistemáticas y una varianza del error conforman la varianza total en dicho modelo teórico. Naturalmente, los resultados prácticos nunca son tan claros. Sin embargo, es notable lo bien que el modelo funciona. Pensar en términos de varianza también es valioso para conceptualizar y analizar los resultados de medición.

Se indica la contribución de cada fuente de varianza. De la varianza total, el 80% es varianza confiable; de la varianza confiable, el factor *A* contribuye con un 30% y el factor *B* contribuye con un 25% y otro 25% es específico de esta prueba. El restante 20% de la varianza total es varianza del error. La prueba se considera bastante confiable, puesto que una proporción importante de la varianza total es confiable o varianza "verdadera". La interpretación de la validez resulta más difícil. Si sólo hubiera un factor, por ejemplo *A*, y contribuyera con el 55% de la varianza total, entonces se podría decir que una proporción considerable de la varianza total sería varianza válida. Se sabría que buena parte de la medición confiable sería la medición de la propiedad conocida como *A*. Ésta sería una afirmación sobre la validez de constructo. Hablando prácticamente, los individuos medidos con la prueba serían ordenados por rangos respecto a *A*, con una confiabilidad adecuada.

No obstante, con el ejemplo hipotético anterior la situación es más compleja. La prueba mide dos factores, *A* y *B*. Podría haber tres conjuntos de órdenes de rango, uno resultante de *A*, uno de *B* y uno *específico*. Mientras que la confiabilidad repetida podría ser alta, si se pensara que se está midiendo únicamente *A*, al grado en que se pensara, la prueba no sería válida. Sin embargo, se podría tener una puntuación para cada individuo, una en *A* y una en *B*. En tal caso la prueba sería válida. Note que aunque se pensara que la prueba está midiendo únicamente *A*, las predicciones con un criterio podrían tener éxito, especialmente si el criterio tuviera mucho de *A* y de *B* en sí mismo. La prueba podría tener validez predictiva aun cuando su validez de constructo fuera cuestionable.

De hecho, los modernos desarrollos en medición indican que tales puntuaciones múltiples han empezado a formar parte, cada vez más, de un procedimiento aceptado.

Relación estadística entre confiabilidad y validez

Aunque aparecen en capítulos diferentes, los temas sobre la confiabilidad y la validez no están separados —ambos tratan con el nivel de excelencia de un instrumento de medición—. En capítulos anteriores se ha visto que es posible tener una medida confiable que no sea válida. Sin embargo, un instrumento de medición sin confiabilidad estaría destinado automáticamente al grupo de los instrumentos "pobres". También se ha mencionado brevemente que si se tiene una medida válida, entonces también se tiene una confiable. En el capítulo 27 se explicó lo que le sucede al coeficiente de confiabilidad cuando se incrementa el tamaño de la prueba. ¿Qué sucede con la validez al incrementarse el tamaño de la prueba? ¿Se ve igualmente afectada que la confiabilidad por el incremento del tamaño? La respuesta contundente es "no". El trabajo clásico de Gullekson (1950) presenta fórmulas para demostrar la relación. Si se añaden suficientes reactivos a la prueba para duplicar el coeficiente de confiabilidad, el coeficiente de validez sólo se incrementa un 41%. Las fórmulas proféticas de la validez por lo general incluyen al coeficiente de confiabilidad de cierta manera y forma. Por ejemplo, existe una fórmula para predecir el coeficiente de validez máximo, con base en el coeficiente de confiabilidad. Con el uso de dicha fórmula es posible obtener un coeficiente de validez más alto que el de confiabilidad. No obstante, en la práctica resulta muy difícil obtener un coeficiente de validez que sea más alto que el de confiabilidad. El razonamiento aquí es que se esperaría que una prueba que se correlaciona consigo misma debería ser mayor que la misma prueba correlacionada con una medida o criterio externo.

Si fuera posible eliminar los errores de medición de la prueba y del criterio, entonces se tendría esencialmente una correlación entre las puntuaciones verdaderas de ambas medidas. Se ha estudiado que los errores de medición tienden a reducir los valores del coeficiente. Es posible, en un sentido hipotético, encontrar cuál podría ser el coeficiente de validez, si se pudiera eliminar el error de medición (i) en el criterio y en la prueba, (ii) sólo en el criterio y (iii) sólo en la prueba. Dichas correcciones son denominadas *correcciones por atenuación*. Si se permite que r_{xy} sea la correlación entre el criterio x y la prueba y, la fórmula para corregir ambas por atenuación es:

$$xy\ corregido\ \ r_{xy} = \frac{r_{xy}}{\sqrt{r_{xx} r_{yy}}}$$

La fórmula para determinar cuál sería la validez si se tuviera un *criterio perfecto* es:

$$r_{\infty y} = \frac{r_{xy}}{\sqrt{r_{xx}}}$$

La fórmula para determinar el coeficiente de validez si se tuviera una *prueba perfecta* es:

$$r_{x\infty} = \frac{r_{xy}}{\sqrt{r_{yy}}}$$

Estas fórmulas no deben utilizarse para tomar decisiones sobre individuos; aunque son útiles para determinar si vale la pena hacer una prueba o un criterio más confiable. Tales fórmulas muestran lo que le sucedería a la validez conforme se hicieran cambios en la confiabilidad.

La validez y confiabilidad de instrumentos de medición psicológicos y educativos

Las mediciones pobres llegan a invalidar cualquier investigación científica. La mayor parte de las críticas a la medición psicológica y educativa, hechas tanto por profesionales como por otras personas, se centra en la validez. Así es como debe ser. Lograr confiabilidad es, en gran parte, un aspecto técnico. Sin embargo, la validez es mucho más que técnica; se centra dentro de la esencia de la propia ciencia. También se centra en la filosofía. La validez de constructo, en particular, tiene un gran sentido filosófico, debido a que se relaciona con la naturaleza de la "realidad" y con la naturaleza de las propiedades que se miden.

A pesar de las dificultades para lograr mediciones psicológicas, sociológicas y educativas válidas y confiables, se ha progresado mucho en este siglo. Existe una creciente comprensión de que todos los instrumentos de medición deben ser examinados crítica y empíricamente, respecto a su confiabilidad y validez. Terminaron los días de tolerancia a la medición inadecuada. Las demandas impuestas por profesionales, las herramientas teóricas y estadísticas disponibles y aquellas que se van desarrollando rápidamente, así como la creciente sofisticación de los estudiantes de posgrado en psicología, sociología y educación, han establecido nuevos estándares más altos que deben ser estimulantes saludables para la imaginación, tanto de los que trabajan en investigación como de quienes desarrollan la medición científica.

RESUMEN DE CAPÍTULO

1. La validez trata con la precisión. ¿El instrumento mide lo que se supone que debe medir?
2. Existen tres tipos de validez
   - de contenido
   - relacionada con el criterio
   - de constructo
3. La validez de contenido se refiere a la adecuación de la representatividad o muestreo del contenido de la prueba.
4. La validez aparente es similar a la validez de contenido, pero no es cuantitativa e incluye una mera inspección visual de la prueba, por parte de revisores sofisticados o no-sofisticados.
5. Existen dos métodos bajo la validez relacionada con el criterio: concurrente y predictiva.
6. La característica distintiva entre la validez concurrente y la predictiva es la relación temporal entre el instrumento y el criterio.
7. Un instrumento con alta validez relacionada con el criterio ayuda a los usuarios de pruebas a tomar mejores decisiones en términos de ubicación, clasificación, selección y evaluación.
8. La validez de constructo busca explicar las diferencias individuales en puntuaciones de pruebas. Trata con conceptos abstractos que pueden contener dos o más dimensiones.
9. La validez de constructo requiere tanto de convergencia como de discriminación.
10. La convergencia establece que los instrumentos que pretendan medir la misma cosa deben estar altamente correlacionados.

11. La discriminación se demuestra cuando instrumentos que se supone miden cosas diferentes tienen una baja correlación.
12. Un método utilizado para demostrar tanto la convergencia como la discriminación es la matriz multirrasgo-multimétodo de Campbell y Fiske (1959).
13. La relación entre la validez y la confiabilidad es susceptible de demostrarse matemáticamente.
14. El conocimiento respecto a la interpretación de las mediciones es importante para los estudios de investigación.
15. Dos temas menos tradicionales respecto a la interpretación y la validez son: la comprobación en referencia al criterio y la comprobación en referencia a la información (o medición con probabilidad admisible).

SUGERENCIAS DE ESTUDIO

1. La literatura sobre la medición es vasta. Las siguientes referencias se eligieron por su excelencia particular o por su relevancia para temas importantes sobre medición. Sin embargo, algunos de los análisis son técnicos y difíciles. El estudiante encontrará análisis elementales sobre confiabilidad y validez en la mayor parte de los libros sobre medición.

Allen, M. J. y Yen, W. M. (1979). *Introduction to measurement theory.* Belmont, California: Brooks/Cole.

Cronbach, L. J. y Meehl, P. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281-302. [Una muy importante contribución a la medición moderna y a la investigación del comportamiento.]

Cureton, E. (1969). Measurement theory, en R. Ebel, V. Noll y R. Bauer (eds.), *Encyclopedia of educational research* (4a. ed.), 785-804. Nueva York: Macmillan. [Un panorama firme y general de la medición, con énfasis en la medición educativa.]

Horst, P. (1966). *Psychological measurement and prediction.* Belmont, California: Wadsworth.

Tryon, R. (1957). Reliability and behavior domain validity: A reformulation and historical critique. *Psychological Bulletin, 54,* 229-249. [Éste es un excelente e importante artículo sobre confiabilidad. Contiene un buen ejemplo trabajado.]

Los siguientes artículos sobre antologías de la medición constituyen fuentes valiosas de los clásicos en el campo. Especialmente los volúmenes de Mehrens y Ebel y de Jackson y Messick.

Anastasi, A. (ed.). (1966). *Testing problems in perspective.* Washington, DC: American Council on Education.

Barnette, W. L. (ed.). (1976). *Reading in psychological tests and measurement* (3a. ed.). Baltimore, MD: Williams y Wilkins.

Chase, C. y Ludlow G. (eds.). (1966). *Readings in educational and psychological measurement.* Boston: Houghton Mifflin.

Jackson, D. y Messick, S. (eds.). (1967). *Problems in human assessment.* Nueva York, McGraw-Hill.

Mehrens, W. y Ebel, R. (eds.). (1967). *Principles of educational and psychological measurement.* Skokie, Illinois: Rand McNally.

2. Un método importante para la validez de estudios es la validez cruzada. Los estudiantes avanzados pueden beneficiarse del ensayo de Mosier en el libro de Chase y Ludlow mencionado anteriormente. Se puede encontrar un breve resumen del ensayo de Mosier en Guilford (1954, p. 406).
3. Los estudiantes más avanzados también querrán saber algo sobre las fijaciones de respuesta —una amenaza para la validez, particularmente para la validez de reactivos e instrumentos de personalidad, actitud y valores—. Las *fijaciones de respuesta* son tendencias a responder los reactivos de ciertas maneras —alto, bajo, aprobar, desaprobar, en extremo, etcétera, independientemente del contenido de los reactivos—. Las puntuaciones resultantes están, por lo tanto, sistemáticamente sesgadas. La literatura es extensa y no puede citarse aquí. Sin embargo, una excelente exposición se encuentra en Nunnally (1978), capítulo 16, especialmente pp. 655 y sig. Los defensores de los efectos de las fijaciones de respuesta en los instrumentos de medición son muy duros en sus afirmaciones. Rorer (1965) ha atacado enfáticamente el tema de las fijaciones de respuesta.

   La posición tomada en este libro es que las fijaciones de respuesta realmente suceden y que en algunas ocasiones tienen efectos considerables, pero que las fuertes declaraciones de los partidarios son exageradas. La mayor parte de la varianza en las medidas bien construidas parece deberse a las variables medidas, y relativamente muy poco a las fijaciones de respuesta. Los investigadores deben estar conscientes de las fijaciones de respuesta y sus posibles efectos negativos sobre los instrumentos de medición, pero no deben tener miedo de utilizar los instrumentos. Si se tomara demasiado en serio a las escuelas de pensamiento sobre las fijaciones de respuesta y sobre lo que se ha llamado el efecto del experimentador (en educación es el efecto Pigmalión) explicado antes, se tendría que abandonar la investigación del comportamiento con excepción, quizás, de la investigación que se realiza con las llamadas medidas no invasivas.
4. Imagine que usted aplicó una prueba con seis reactivos a seis personas. Las puntuaciones de cada reactivo de cada persona se presentan abajo. Suponga que también aplicó otra prueba con seis reactivos a otras seis personas. Las puntuaciones también se incluyen abajo. Las puntuaciones de la primera prueba, I, se presentan a la izquierda; las puntuaciones de la segunda prueba, II, se presentan a la derecha.

I							II						
	Reactivos							*Reactivos*					
Personas	*a*	*b*	*c*	*d*	*e*	*f*	Personas	*a*	*b*	*c*	*d*	*e*	*f*
1	6	6	7	5	6	5	1	6	4	5	6	6	3
2	6	4	5	5	4	5	2	6	2	7	4	4	4
3	5	4	7	6	4	3	3	5	6	5	3	4	2
4	3	2	5	3	4	4	4	3	4	4	5	4	5
5	2	3	4	4	3	2	5	2	1	7	1	3	5
6	2	1	3	1	0	2	6	2	3	3	5	0	2

Las puntuaciones en II son las mismas que en I, excepto que el orden de las puntuaciones de los reactivos (*b*), (*c*), (*d*) y (*f*) se ha cambiado.

   a) Realice un análisis de varianza de dos factores con cada uno de los conjuntos de puntuaciones. Compare e interprete las razones *F*. Ponga especial atención a la razón *F* para *Personas* (individuos).

b) Calcule $r_{tt} = (V_{ind} - V_e)/ V_{ind}$ para I y II. Interprete las dos r_{tt}. ¿Por qué son tan diferentes?

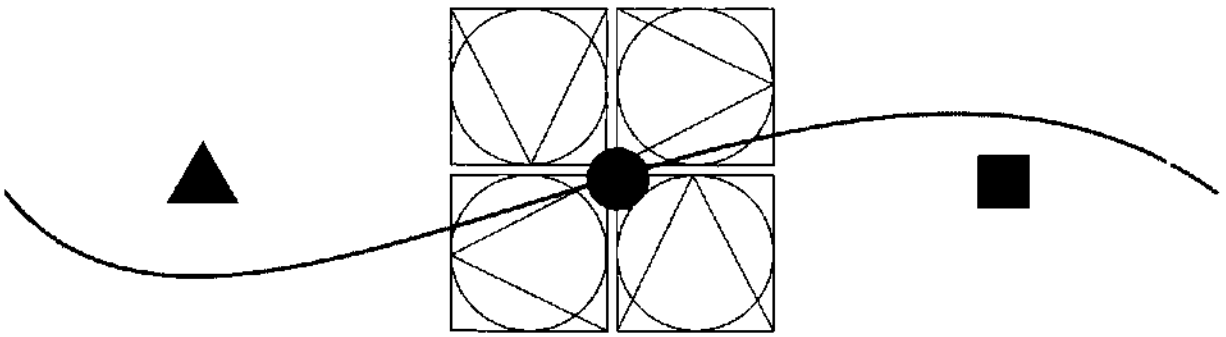
c) Sume los reactivos impares a través de los renglones; sume los reactivos pares. Compare los órdenes de rango y los rangos de los totales impares, de los totales pares y de los totales de los seis reactivos. Los coeficientes de correlación entre los reactivos impares y pares, corregidos, son .98 y .30. Explique por qué son tan diferentes. ¿Qué significan?

d) Suponga que había 100 personas y 60 reactivos. ¿Habría cambiado esto los procedimientos y el razonamiento subyacente? ¿Habría afectado, el efecto de cambiar el orden de, por ejemplo, cinco a diez reactivos, a las r_{tt} tanto como en estos ejemplos? Si no fuese así, ¿por qué no?

[Respuestas: ***a***) I: $F_{reactivos} = 3.79$ (.05); $F_{personas} = 20.44$ (.001). II: $F_{reactivos} = 1.03$ (n.s); $F_{personas} = 1.91$ (n.s). ***b***) I: $r_{tt} = .95$; II: $r_{tt} = .48$.]

Parte Nueve

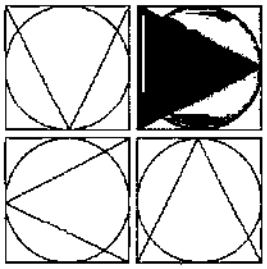
Métodos de observación y de recolección de datos



Capítulo 29
Entrevistas e inventarios de entrevistas

Capítulo 30
Pruebas y escalas objetivas

Capítulo 31
Observaciones del comportamiento y sociometría



Capítulo 29

Entrevistas e inventarios de entrevistas

La entrevista es quizás la técnica de uso más frecuente para obtener información de la gente. Ha sido y aún es utilizada en todo tipo de situaciones prácticas: el abogado obtiene información de un cliente; el médico conoce sobre el paciente; el oficial de admisiones o el profesor determina la adecuación de estudiantes a determinadas escuelas, departamentos y programas. Sin embargo, sólo hasta hace poco se ha utilizado la entrevista de forma sistemática para propósitos científicos, tanto en el laboratorio como en el campo.

Los métodos de recolección de datos se clasifican de acuerdo a qué tan directos son. Si se desea saber algo sobre las personas, se les puede preguntar directamente. Ellos ofrecen o no una respuesta. Por otro lado, es posible preguntar de forma indirecta. Se puede utilizar un estímulo ambiguo como una fotografía borrosa, una mancha de tinta o una pregunta vaga, y después preguntar respecto a las impresiones de los estímulos, bajo el supuesto de que los entrevistados darán la información requerida sin saber que lo están haciendo. Esta técnica es bastante indirecta. La mayor parte de los métodos de recolección de datos utilizados en la investigación psicológica y sociológica son relativamente directos o moderadamente indirectos. En pocas ocasiones se utilizan medios muy indirectos.

Las entrevistas y los inventarios (cuestionarios) por lo general son bastante directos, lo cual representa tanto una fortaleza como una debilidad. Tienen fortaleza porque gran

cantidad de la información requerida en la investigación social científica se obtiene de los entrevistados por medio de preguntas directas. Aunque las preguntas deben manejarse con sumo cuidado, los entrevistados pueden, y generalmente lo hacen, dar mucha información de forma directa. No obstante, existe información de naturaleza más difícil que los entrevistados quizá no estén dispuestos a dar fácil y directamente —por ejemplo, información sobre sus ingresos, relaciones sexuales y algunas actitudes hacia la religión o hacia los grupos minoritarios—. En tales casos, las preguntas directas llegan a generar datos que no son válidos. Sin embargo, si se manejan en forma apropiada, aun el material personal o polémico puede obtenerse exitosamente por medio de entrevistas e inventarios.

La entrevista es probablemente uno de los métodos más antiguos y más utilizados para conseguir información. Posee importantes cualidades que las pruebas y escalas objetivas y las observaciones del comportamiento no tienen. Una entrevista puede proporcionar una gran cantidad de información si se utiliza con un inventario bien realizado. Es flexible y se adapta a situaciones individuales, y puede usarse con frecuencia cuando ningún otro método es posible o adecuado. Estas cualidades la hacen especialmente adecuada para la investigación con niños. Los métodos y consideraciones sobre las entrevistas con niños pueden encontrarse en Aldridge y Wood (1998) y en Poole y Lamb (1998). Minkes, Robinson y Weston (1994) ofrecen explicaciones sobre la forma de entrevistar a niños con discapacidades. Ellis (1989) describe cómo conducir una entrevista con niños superdotados canadienses. Si un entrevistador sabe que el entrevistado, especialmente un niño, no entiende una pregunta, puede, dentro de ciertos límites, repetir o replantear la pregunta. Las preguntas sobre deseos, aspiraciones y ansiedades pueden plantearse de tal manera que produzcan información precisa. De mayor importancia, quizás, es el hecho de que la entrevista permite explorar el contexto y las razones de las respuestas a las preguntas. McReynolds (1989) sintetiza el estado de los instrumentos de medición clínica, de los cuales uno es el *inventario de entrevista*.

La mayor desventaja de la entrevista y de su inventario acompañante es de índole práctica. Las entrevistas toman mucho tiempo. El obtener información de un individuo llega a tomar tanto como una hora e incluso dos horas. Una gran inversión de tiempo implica esfuerzo y dinero. Andrews (1974) determina los requisitos de un estudio de investigación bien conducido, que utilice la entrevista, en términos del reclutamiento, entrenamiento, selección y supervisión. Uno de los componentes importantes de la entrevista es la supervisión. Andrews menciona por lo menos nueve responsabilidades que un supervisor debe tener. Por lo tanto, siempre que una técnica más económica responda a los propósitos de investigación, no deben utilizarse las entrevistas. Emory (1976) cita investigación realizada respecto a las características del entrevistador. Tales estudios hallaron evidencia de que características triviales podían influir en los resultados de la entrevista. Por ejemplo, Emory cita el hecho de que las mujeres son mejores entrevistadores que los hombres, y que los hombres casados son mejores que las mujeres solteras. También cita un estudio realizado por el Centro Nacional de Investigación de Opinión (National Opinion Research Center, NORC) que relaciona ciertas características con la calidad de la entrevista. El entrenar entrevistadores para que produzcan el mismo nivel de calidad requiere de tiempo, recursos e inclusive quizás de experiencia previa.

Las entrevistas e inventarios como herramientas de la ciencia

Las entrevistas e inventarios han sido utilizados, en su mayor parte, simplemente para reunir los llamados hechos. El uso más importante de la entrevista debe ser el estudio de