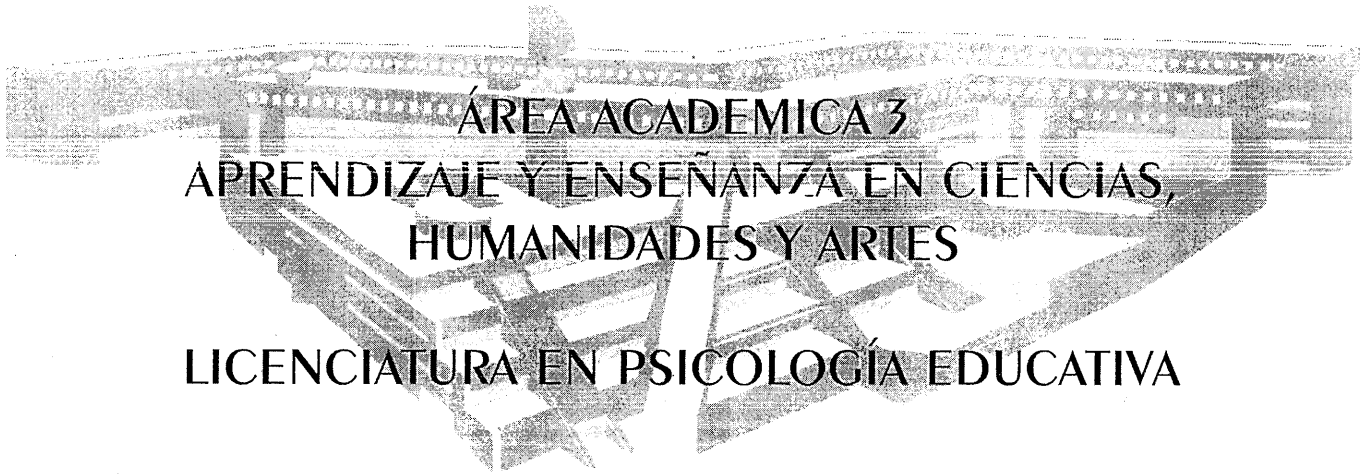
UNIVERSIDAD PEDAGÓGICA NACIONAL
AJUSCO



ÁREA ACADEMICA 3
APRENDIZAJE Y ENSEÑANZA EN CIENCIAS, HUMANIDADES Y ARTES

LICENCIATURA EN PSICOLOGÍA EDUCATIVA

Métodos Cuantitativos

Programa del curso
6 horas, 10 créditos

Quinto semestre, Plan 2009

Sección de Reprografía.

Presentación

La asignatura de Métodos Cuantitativos está dentro de la Línea Metodológica de la licenciatura en Psicología Educativa. Las materias que le anteceden son Estadística I (2° semestre), Estadística II (3er semestre) y Enfoques y Métodos de Investigación (4° semestre); estas materias les brindan a los alumnos los conocimientos y habilidades relativos a los tipos de investigación con que se puede abordar el estudio de fenómenos sociales, en particular los que atañen a la educación, así como a las herramientas estadísticas para el análisis de datos surgidos de ellas. A partir de esta asignatura, los estudiantes irán adquiriendo herramientas para recabar información relativa a los procesos educativos: primero los de índole cuantitativa, a los que está dirigida esta materia, y posteriormente los de índole cualitativa, que se cubrirán en el sexto semestre. Cierran la Línea Metodológica dos seminarios, uno de Diseño Metodológico, y uno de Titulación, respectivamente de 7° y 8° semestres, en los que los alumnos podrán integrar todos los conocimientos adquiridos y diseñar investigaciones.

Toda la Línea Metodológica, y muy en particular la asignatura de Métodos Cuantitativos, pretende apoyar dos aspectos fundamentales del perfil de egreso de la Licenciatura: "Atender e investigar problemas de la enseñanza escolarizada en el Sistema Educativo Nacional" y "Establecer y desarrollar procedimientos para atender e investigar problemas de aprendizaje escolar". La manera en que esta asignatura apoya dichas metas es proporcionándole al alumno una visión general sobre los procesos de medición, en particular en Psicología Educativa, y un manejo práctico de algunos instrumentos con los que se realiza la medición y la recolección de la información correspondiente.

Es por ello que el objetivo general de la asignatura es brindar a los estudiantes los conocimientos teórico-metodológicos, así como las actividades para desarrollar habilidades que les permitan identificar: cómo los métodos cuantitativos usan la medición para obtener información, qué tipo de métodos hay, cómo se aplican, y cómo se analiza la información recabada - para lo cual podrán aplicar lo aprendido en los cursos anteriores.

El enfoque que se le da a la materia se centra, principalmente, en el *aprender a hacer*, es decir, en el conocimiento de tipo procedimental, de tal modo que los estudiantes adquieran la habilidad de identificar, seleccionar y aplicar el instrumento

adecuado, de entre los tres tipos de instrumentos de corte cuantitativo más usuales en Psicología Educativa -observación, cuestionarios y escalas-, y que analicen la información surgida de dicha aplicación.

De esta manera, el curso se desarrolla en tres fases, que corresponden a cada una de las tres Unidades del programa. En una primera fase se abordan los conceptos generales de la medición; la segunda se dirige a la aplicación; y la tercera al análisis y reporte de los resultados. Cabe destacar que la asignatura da un énfasis importante al trabajo de campo, y por ende a la segunda fase: en ella se busca que los estudiantes puedan experimentar, en los escenarios naturales, la situación de aplicar instrumentos con muestras de sujetos inmersos en un contexto educativo. Para ello se propone que los instrumentos a aplicar estén previamente construidos, sin dejar de lado una visión muy general de los complejos procesos implicados en la construcción. El hecho de utilizar instrumentos previamente diseñados permite optimizar tiempo y esfuerzo en beneficio del aprendizaje de los contenidos de la materia.

Propósito general

Al finalizar el curso, los estudiantes habrán adquirido los conocimientos teórico-metodológicos y desarrollado habilidades a través de actividades que les permitan identificar: cómo los métodos cuantitativos usan la medición para obtener información, qué tipo de métodos hay, cómo se aplican, y cómo se reporta la información recabada.

UNIDAD 1. ELEMENTOS DE MEDICIÓN

Propósito de la unidad

Que los alumnos tengan una visión general de lo que significa y lo que implica un proceso de medición, así como de conceptos relacionados con ella, tales como niveles de medición, validez y confiabilidad.

Temas

1. Concepto de medición
2. La medición en psicología
3. Niveles de medición
4. Validez y confiabilidad de instrumentos de medición

ACTIVIDADES

Actividad introductorias 1

Lectura de la bibliografía y discusión dirigida

Recursos

- Antología
- Pizarrón, plumones

Duración: 18 horas

Bibliografía básica

De la Florida R., M. A. (2004). ¿Quién inventó el metro? *Algarabía,* 7(12), 58-60

Kerlinger, F. N. (2002). *Investigación del comportamiento, métodos de investigación en ciencias sociales* (4ª edición). México: Mc Graw-Hill, capítulos 26, 27 y 28.

Müller, M., Ballesteros, S., Bernal, M.A., Bonilla, B.J., Escalona, J., González R.A., Luna K. y Bernal. U.M.I. (2006). Medir para vivir. *¿Cómo ves?, 8*(87), 16-19.

Bibliografía complementaria

Hernández, R.; Fernández, C. y Baptista, P. (2007). *Fundamentos de metodología de la investigación educativa*. Madrid: McGraw Hill.

Kerlinger, F. N. (2002). *Investigación del comportamiento, métodos de investigación en ciencias sociales* (4ª edición) , capítulo 3. México: Mc Graw-Hill.

William Travers, R. M. (1979). *Introducción a la investigación educacional.* Buenos Aires: Paidos.

UNIDAD 2. TRES INSTRUMENTOS DE MEDICIÓN EN PSICOLOGÍA Y EDUCACIÓN

Propósito de la unidad

Que los alumnos conozcan las características de cada uno de tres tipos de instrumentos, y que los apliquen a sujetos dentro de una comunidad educativa.

Temas

1. **Observación**
    - Definición / características
    - Tipos de observación cuantitativa (Semi-estructurada, Estructurada, Categorial, No estructurada)
    - Diferencias con la observación participante
    - Guías de observación
    - Aplicación
2. **Cuestionarios**
    - Definición / características
    - Tipos de preguntas; preguntas con respuestas únicas y no únicas
    - Aplicación

3. **Escalas**

    - Definición / características
    - Tipos de escalas
    - Escalamiento
    - Aplicación

ACTIVIDADES

- ***Actividades 2 (observación), 3 (cuestionarios) y 4 (escalas)***
- Aplicar un instrumento de observación, un cuestionario y una escala

NOTA: En el material proporcionado, se proponen dos instrumentos de observación, dos cuestionarios y dos escalas, para elegir uno de cada tipo. Se sugiere que en cada grupo se aplique al menos un instrumento específicamente relacionado con la Psicología y al menos uno **no** específicamente relacionado con la Psicología, con la finalidad de que los estudiantes tengan una visión general del amplio rango de circunstancias en que estos instrumentos se aplican.

Recursos

- Antología
- Pizarrón, plumones
- CD de actividades en computadora
- Guías, cuestionarios, etc.

Duración: 60 horas; se sugiere consagrar 12 al primer tema y 24 a cada uno de los otros dos.

Bibliografía básica

Tema 1:

Espinosa Aramburu, M. C. (1997). Metodología Observacional. México: Facultad de psicología UNAM.

Tema 2:

Gómez, A. M. (2007). La investigación educativa: claves teóricas. España: MacGraw Hill, pp. 99-135.

Reyes L., I. y García B., L.F. (2010). Procedimiento de validación psicométrica culturalmente relevante: un ejemplo. *La psicología social en México*, 13, 625-630.

Tema 3:

Padua, J. (1979). *Técnicas de investigación aplicadas a las ciencias sociales.* Cap. 6. México: El Colegio de México/FCE.

Bibliografía complementaria

Kerlinger, F. N. (2002). *Investigación del comportamiento, métodos de investigación en ciencias sociales* (4ª edición) , capítulo 30. México: Mc Graw-Hill.

UNIDAD 3. ANÁLISIS Y REPORTE DE LA INFORMACIÓN RECABADA

Propósito de la unidad

Que los alumnos realicen el análisis de la información recabada durante la aplicación de alguno de los instrumentos utilizados (observación, cuestionario o escala) y comprendan la importancia de difundir entre *la comunidad científica, académica o profesional*, los hallazgos derivados de los estudios realizados.

Tema único: Informe de resultados

ACTIVIDADES

Seleccionar un de las aplicaciones realizadas en el desarrollo de la Unidad II.

Elaborar un documento escrito en el que se informe acerca del trabajo realizado, considerando los elementos arriba listados.

Realizar una presentación oral (apoyada en presentación electrónica) del informe elaborado

Actividad 5: guía para el contenido del reporte

Asistir a asesoría

Recursos

- Pizarrón, plumones, equipo de informática y proyector

Duración: 18 horas

Bibliografía básica

American Psychological Association (2002). Manual de estilo de publicaciones de la American Psychological Association. México: Manual Moderno. Cap. 155-198; 299-314.

Bibliografía complementaria

Padua, J. (1979). *Técnicas de investigación aplicadas a las ciencias sociales.* Cap. 9. México: El Colegio de México/FCE.

UPN. Programa y antologías de Estadística I y II.

Programa elaborado por:

Mtra. Magdalena Aguirre Tobón

Dra. Silvia Alatorre Frenk

Lic. David Díaz Mercado

Dr. Jorge García Villanueva

Mtro. Cuauhtémoc Gerardo Pérez López

Dr. Armando Ruiz Badillo

JUNIO 2011

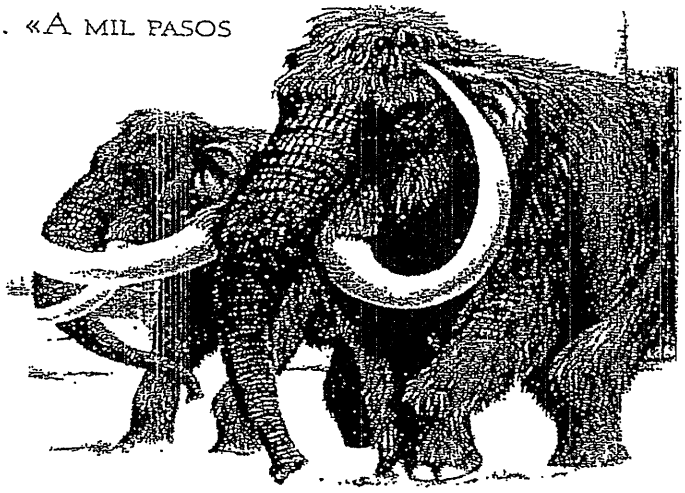
SERENDIPIA

De la Florida Rivero, M. A. (2004). ¿Quién inventó el metro? *Algarabía* , 58-60.

¿Quién inventó el metro?

por Manuel Alonso de la Florida Rivero

LA IMPORTANCIA DE TENER PARÁMETROS DE MEDICIÓN ES TAN ANTIGUA COMO EL HOMBRE MISMO, PERO ¿CUÁNTO MEDÍA EL MAMUT?, «PUES COMO TRES CHUCHOS» DIJO EL CAZADOR CAVERNÍCOLA. «¿Y DÓNDE LO ENCONTRASTE?». «A MIL PASOS DE LA PIEDRA ROJA».



El pie, la mano, el pulgar, la vara, son los primeros instrumentos de medición que usó el hombre. Pero estas unidades tenían un gran problema: que eran y son arbitrarias, imprecisas, cambian de un lugar a otro y lo que es peor, de una persona a otra: mi pie y mi pulgar miden distinto que los de mi hermano. De ahí surge la necesidad de crear acuerdos para fijar lo que hoy conocemos como sistema de medidas.

Uno de los primeros que intentó establecer un sistema único de medidas fue Carlomagno, quien trató de imponer las unidades «Parisinas» en toda Francia. Fracasó porque los nobles y los comerciantes se beneficiaban con la confusión en las conversiones de una medida a otra.

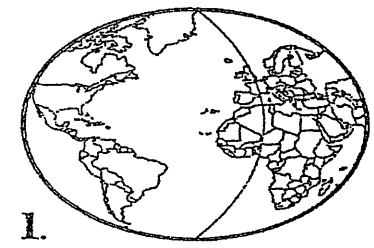
Pero la gran idea, la idea genial, surge en la época de la Revolución Francesa cuando la Asamblea Nacional le pide a la Academia Francesa de las Ciencias que forme una comisión de científicos para que desarrollen el sistema de medidas de la nueva república, el cual debía estar basado en dos principios fundamentales: la observación científica y la fracción decimal.

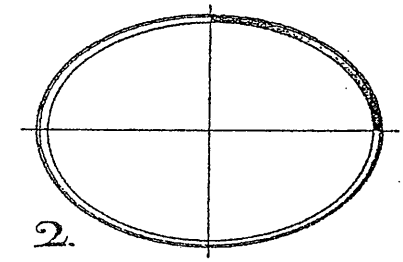
De esta manera surge el Sistema Métrico Decimal que utilizamos hoy en día. La unidad de longitud fue definida

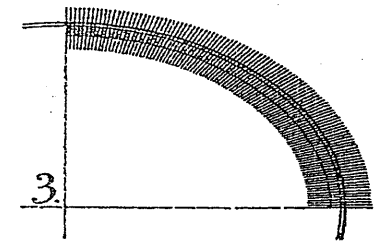
por J. L. Lagrange y Pierre Simon Laplace, entre otros, como una diezmillonésima de la cuarta parte del meridiano terrestre que pasa por París, esta definición fue aceptada por la Asamblea Nacional el 26 de marzo de 1791 y recibió el nombre de *metro* en 1793.

¿Pero qué quiere decir 1/10'000,000 del meridiano terrestre que pasa por París? Entendámoslo en los siguientes pasos:

1. Trace una línea imaginaria sobre un globo terráqueo que pase por los polos y por París: obtuvo el perímetro o meridiano de la Tierra.
2. Divida esta línea en cuatro partes y tendrá la longitud del cuadrante.
3. Ahora divida el cuadrante entre diez millones y obtendrá un metro.


1.


2.


3.

La definición del metro reflejaba el gran interés de los científicos franceses en la figura de la Tierra.

El sistema métrico se adoptó oficialmente en Francia el 7 de abril de 1795 y se formalizó mediante un decreto donde se adoptaron las definiciones que usamos ahora. Para contar con un metro modelo se construyó una barra de metal con las medidas que se tomaron a partir de las mediciones de arco de meridiano que hizo Nicolas Lacaille en 1740.

En 1799, en una conferencia científica celebrada en Francia en la que hubo representantes de los Países Bajos, España, Dinamarca e Italia, se validaron los cálculos para diseñar los prototipos modelo y construir los patrones permanentes; el del metro se fabricó en platino y fue depositado en los Archivos de la República ese mismo año.

El 20 de mayo de 1875 los delegados de 17 países —incluido Estados Unidos— firmaron en París el Tratado

del metro» en el que se designa la Oficina Internacional de Pesas y Medidas que se establece en el Pavillon de Breteuil cerca de París. En 1889 le son encargadas a una casa inglesa las barras que servirían de patrón internacional al metro, mismas que son fabricadas de una aleación de 90% platino y 10% iridio, tomando como medida el metro francés original.

El metro patrón de 1889 fue verificado y en 1892 fue abandonada la referencia al meridiano terrestre.

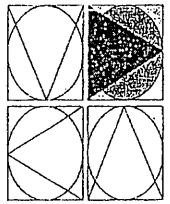
Para redefinir el metro con mayor precisión se han ido escogiendo distintas fuentes de luz, como la del kriptón-86 roja-naranja, cuya longitud de onda fue obtenida por comparación directa con la longitud de onda del cadmio. En 1960, el XI Consejo General de Pesas y Medidas definió el metro como la longitud igual a 1'650,763.73 longitudes de onda de la línea espectral del kriptón-86 roja-naranja.

La definición válida para el metro fue determinada en 1983 por el XVII Consejo General de Pesas y Medidas —CGPM— como «la longitud del camino atravesado por la luz en el vacío durante un intervalo de tiempo de 1/299'792,458 de un segundo». Esta definición fija la velocidad de la luz en 299'792,458 m/s exactamente.

Para concluir, quisiera decir que el símbolo de metro es m, no *M*, ni *mt.*, ni *mts.*, ni ninguna de esas abreviaturas. No se hagan bolas.

UNIVERSIDAD
PEDAGÓGICA
NACIONAL

Kelinger, F. N., & Lee, H. B. (2001). *Investigación del comportamiento.* México: Mc Graw Hill. Capítulo 26, 27 y 28. Pp.565-625



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.

Este ejemplo quizá parezca un poco ridículo; pero la mayoría de nosotros atravesamos, en gran parte, por el mismo procedimiento durante toda nuestra vida. Con frecuencia juzgamos qué tan "simpática", qué tan "fuerte", qué tan "inteligente" es la gente, tan sólo con verla y hablar con ella. Únicamente parece insensato cuando se presenta como un ejemplo serio de la medición. Insensato o serio, *es* un ejemplo de medición, ya que satisface la definición. El juez asignó valores numéricos a "objetos" de acuerdo con reglas. Los objetos, los valores numéricos y las reglas para asignar los valores numéricos a los objetos estaban especificados. Los valores numéricos fueron 1, 2, 3, 4 y 5; los objetos eran los estudiantes; las reglas para la asignación de los valores numéricos a los objetos estaban contenidas en las instrucciones dadas al juez. El producto final del trabajo —los valores numéricos— podía utilizarse después para calcular medidas de relación, análisis de varianza y otros aspectos similares.

La definición de medición no incluye estipulaciones acerca de la calidad del procedimiento de medición. Tan sólo dice que, de alguna manera, se asignen valores numéricos a objetos o eventos. El "de alguna manera" es naturalmente importante —pero no para la definición—. La medición es un juego que se practica con objetos y valores numéricos; los juegos tienen reglas. Por supuesto es importante, por otras razones, aparte de que las reglas sean "buenas"; pero aunque las reglas sean "buenas" o "malas", el procedimiento continúa siendo de medición. ¿Por qué el énfasis en la definición de medición y en su calidad de "reglamentada"? Existen tres razones.

En primer lugar, la medición, en especial la medición en psicología y en educación, está mal entendida. No es difícil entender ciertas mediciones utilizadas en las ciencias naturales —por ejemplo, duración, peso y volumen—. Inclusive las medidas que se alejan más del sentido común pueden comprenderse sin deformar demasiado ciertos conceptos elementales e intuitivos. Sin embargo, es más difícil entender que la medición de las características de individuos y grupos, tales como inteligencia, agresividad, cohesión y ansiedad, implica *básica y esencialmente* el mismo razonamiento y procedimiento general. De hecho, muchos consideran que no es posible hacerlo. Entonces, saber y entender que la medición es la asignación de valores numéricos a objetos o eventos por medio de reglas, ayuda a borrar conceptos erróneos y confusos sobre la medición en psicología y educación.

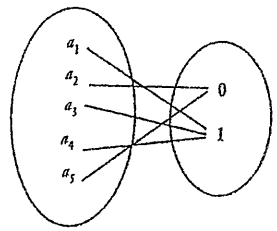
En segundo lugar, la definición indica que, si las reglas pueden ser establecidas con una base racional o empírica, entonces la medición de cualquier cosa es *teóricamente* posible, lo cual amplía en gran medida los horizontes de medición del científico. El científico no rechazará la posibilidad de medir alguna propiedad debido a que ésta sea compleja y difícil de alcanzar. Establece que la medición es un juego que puede jugarse o no con esta o aquella propiedad, en un momento determinado. Nunca se rechazará participar en el juego de medición, aunque el científico comprenda las dificultades que implica.

En tercer lugar, la definición alerta sobre el foco neutral y esencial de la medición y de los procedimientos de medición, y sobre la necesidad del establecimiento de "buenas" reglas, reglas cuya virtud pueda ser probada empíricamente. Un procedimiento de medición no es mejor que sus reglas. Las reglas dadas en el ejemplo anterior eran pobres. El procedimiento era un procedimiento de medición; se satisfizo la definición. Sin embargo, fue un procedimiento pobre por razones que se van a aclarar más adelante.

Definición de medición

Se repite la definición: "la medición es la asignación de valores numéricos a objetos y eventos, de acuerdo con ciertas reglas". Un *valor numérico* es un símbolo con la forma 1, 2, 3,..., o I, II, III. No posee ningún significado cuantitativo a menos que éste le sea asignado;

FIGURA 26.1



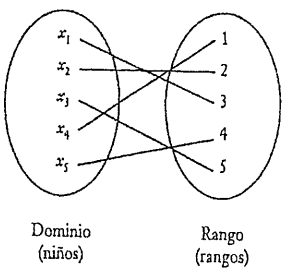
se trata simplemente de un símbolo de un tipo especial. Puede utilizarse para marcar objetos, como jugadores de beisbol, bolas de billar o individuos, seleccionados en una muestra, a partir de un universo. También podría utilizarse el término *símbolo* en la definición. Es posible, e inclusive necesario, asignar símbolos a objetos, o conjuntos de objetos, de acuerdo con reglas. El término *valor numérico* se utiliza porque la medición generalmente usa valores numéricos, los cuales se convierten en números después de que se les ha asignado un significado cuantitativo. Por lo tanto, un *número* es un valor numérico al que se le ha asignado un significado cuantitativo.

El término "asignado" en la definición significa *representación*. Recuerde que anteriormente se habló de representar los objetos de un conjunto sobre los objetos de otro conjunto. Una función, *f*, es una regla de *correspondencia*. Es una regla que asigna a cada miembro de un conjunto algún miembro de otro conjunto. Los miembros de los dos conjuntos pueden ser *cualquier* objeto. En matemáticas, los miembros por lo común son números y símbolos algebraicos. En investigación, los miembros de un conjunto pueden ser individuos o símbolos que representen individuos, y los miembros del otro conjunto pueden ser valores numéricos o números. En la mayor parte de la medición en psicología y educación, los valores numéricos y los números son representados o asignados a individuos. En la representación casi siempre los miembros del dominio se consideran representados sobre los miembros del rango. Para mantener consistencia con la anterior definición de medición, y para poder concebir siempre el proceso de medición como una función, se ha invertido la representación. Esta concepción de la representación también es consistente con la definición previa de una función como una regla que asigna a cada miembro del dominio de un conjunto, un miembro del rango. La regla describe cómo se ordenan los pares.

El trabajo más interesante —y más difícil— de la medición es la regla. Una regla es una guía, un método, una orden que indica qué hacer. Una regla matemática es *f*, una función; *f* es una regla para asignar objetos de un conjunto a los objetos de otro conjunto. En medición una regla podría decir: "asigne los valores numéricos del 1 al 5 a individuos, de acuerdo con qué tan agradables sean. Si un individuo es muy agradable, asígnele el número 5. Si un individuo no es agradable, asígnele el número 1. Asigne a los individuos entre estos límites números entre los límites". Otra regla ya se ha utilizado varias veces: "si una persona es mujer, asígnele un 1; si es hombre, asígnele un 0". Por supuesto que de antemano debe tenerse una regla o conjunto de reglas que definan a hombre y mujer.

Suponga que se tiene un conjunto, *A*, de cinco personas, tres mujeres y dos hombres: a_1, a_3 y a_4 son mujeres; a_2 y a_5 son hombres. Se desea medir la variable *sexo*. Suponiendo que se tiene una regla previa que permite determinar el sexo de manera precisa, se utiliza

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"

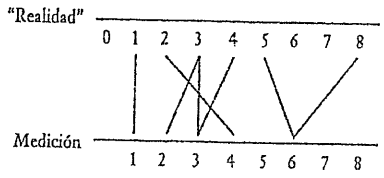
Como se ha visto, la medición puede ser un asunto sin sentido. ¿Cómo evitar esto? La definición de conjuntos de objetos a medir, la definición de conjuntos numéricos a partir de los cuales se asignan valores numéricos a los objetos que se miden, y las reglas de asignación o correspondencia, deben ligarse 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,

FIGURA 26.3



que el ser omnisciente sabía que los valores "verdaderos" de persistencia van de 0 a 8, mientras que su sistema de medición sólo abarca del 1 al 7.

Aunque este ejemplo es un poco extravagante, presenta de forma ordinaria la naturaleza del problema del isomorfismo. La pregunta fundamental que debe plantearse respecto a cualquier procedimiento de medición es: ¿el proceso de medición es isomórfico a la realidad? Usted no estuvo muy lejos al medir la persistencia; el problema es que en pocas ocasiones se descubre de esta forma el grado de correspondencia de las mediciones con la "realidad". De hecho, ¡con frecuencia ni siquiera se sabe si se está midiendo lo que se intenta medir! A pesar de esa dificultad, los científicos deben medir, de alguna forma, el isomorfismo que tienen los juegos de números de medición que ellos juegan, con la "realidad".

Propiedades, constructos e indicadores de objetos

Se dice que se miden objetos, pero esto no es totalmente verdad: se miden sus propiedades o sus características. Sin embargo, inclusive esta calificación no es del todo cierta. En realidad se miden *indicadores* de las propiedades de los objetos, así que cuando se afirma que se miden objetos, en realidad se está diciendo que se miden indicadores de las propiedades de los objetos, lo cual ocurre, por lo general, en toda la ciencia; aunque las propiedades de algunos objetos naturales están más cerca de la observación directa que otras. Por ejemplo, la propiedad del sexo asociada con objetos animales está directamente relacionada con la observación directa. En cuanto las propiedades físicas relativamente simples se dejan atrás por propiedades más complejas y evasivas —las cuales son de mucho mayor interés para los científicos sociales y para los educadores—, la observación directa de las propiedades se vuelve imposible. La hostilidad no puede observarse de forma directa, tampoco la moral, la ansiedad, la inteligencia, la creatividad ni el talento. Dichas propiedades o características deben siempre *inferirse* a partir de la observación de supuestos indicadores de dichas propiedades.

Indicador es simplemente un término conveniente utilizado para significar algo que apunta hacia algo más. Si un niño les pega continuamente a otros niños, se afirma que su conducta es un indicador de su hostilidad subyacente. Si las manos de alguien sudan excesivamente, se dice que la persona está ansiosa. Si una niña toca hermosamente una pieza de Schubert de forma improvisada, se afirma que ella tiene "talento". Si un niño marca correctamente un determinado número de reactivos de tipo objetivo, en una prueba de rendimiento, se dice que tiene cierto nivel de rendimiento. En cada uno de estos casos, alguna conducta identificable constituye un indicador de una propiedad subyacente. En efecto, se tiene una base menos sólida cuando se hacen dichas inferencias a partir de la

conducta observada, que cuando se observan directamente propiedades tales como el color del cabello, tamaño corporal y sexo. Medir el grado de cooperación, dependencia e imaginación de un niño es muy diferente que medir la estatura el peso o el desarrollo del hueso de la muñeca de un niño. El proceso fundamental de medición es el mismo; pero las reglas son mucho más difíciles de prescribir. Además, las observaciones de las propiedades psicológicas están mucho más alejadas de las propiedades reales, que las de las propiedades físicas. Ésta es, quizá, la mayor dificultad de la medición en psicología y educación.

Los indicadores, a partir de los cuales se infieren propiedades, se especifican por medio de definiciones operacionales, las cuales especifican las actividades u "operaciones" necesarias para medir variables o constructos. Un *constructo* es un nombre inventado para una propiedad. Muchos constructos se han utilizado en capítulos previos: autoritarismo, rendimiento, clase social, inteligencia, persistencia, etcétera. Los conceptos o constructos en discusión también se denominan "variables latentes". Ésta es una expresión importante que se está utilizando con éxito en lo que se llama análisis estructural de covarianza, o análisis causal. Una *variable latente* es un constructo, una variable no observable, que se supone subyace a diversas conductas, y que se utiliza para "explicar" dichas conductas. Por ejemplo, "habilidad verbal", "conservadurismo" y "ansiedad" son variables latentes. Su uso se explicará más adelante en el libro, cuando se estudie el análisis factorial y el análisis estructural de covarianza.

Los constructos, comúnmente llamados, de manera algo imprecisa, variables, se definen de dos formas generales en la ciencia: por medio de otros constructos, y por medio de procedimientos experimentales y de medición. Éstas fueron llamadas anteriormente *definiciones operacionales y constitutivas*. Es necesaria una definición operacional para medir una propiedad o un constructo. Ello se hace especificando las observaciones de los indicadores conductuales de las propiedades.

Los valores numéricos se asignan a los indicadores conductuales de propiedades; entonces, después de realizar las observaciones de los indicadores, los números (valores numéricos) son sustituidos por los indicadores y luego se analizan estadísticamente. Como ejemplo, considere a investigadores que están trabajando con la relación entre inteligencia y honestidad. Ellos definen operacionalmente la *inteligencia* como las puntuaciones en una prueba de inteligencia. La *honestidad* se define operacionalmente como las observaciones en una situación artificial en donde se permite hacer trampa o no a unos alumnos. Los valores numéricos de inteligencia, asignados a los alumnos, pueden ser el número total de reactivos correctos en la prueba, o alguna otra forma de puntuación. Los valores numéricos de honestidad asignados a los alumnos son el número de veces que no hicieron trampa cuando pudieron haberla hecho. Los dos conjuntos de números pueden correlacionarse o analizarse de alguna otra manera. El coeficiente de correlación es, por ejemplo, .55, que resulta significativo al nivel .01. Todo esto es bastante directo y familiar. Lo que no es tan directo ni familiar es lo siguiente: si los investigadores llegan a la conclusión de que existe una relación positiva significativa entre la inteligencia y la honestidad, ellos están dando un gran salto inferencial, desde indicadores conductuales en la forma de marcas sobre un papel y desde observaciones de la conducta sobre "hacer trampa", hasta propiedades psicológicas. Debe resultar bastante obvio que quizá ellos estén equivocados.

Niveles de medición y escalación

Los niveles de medición, las escalas asociadas con los niveles y los estadísticos apropiados para los niveles constituyen problemas complejos, e inclusive controvertidos. Las

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 esignados, es posible contar a los miembros de los conjuntos. En caso de dicotomía, la egla de conteo fue expresada en el capítulo 4. Si un miembro de U posee la característica n cuestión, por ejemplo, condición masculina, entonces se asigna 1. Si el miembro no osee la característica, entonces se asigna 0 (véase figura 26.1). Cuando los miembros del onjunto se cuentan de esta manera, todos los objetos de un subconjunto se consideran uales 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 ema aceptan únicamente la medición ordinal, de intervalo y de razón; mientras que otros firman 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.

Medición ordinal

La medición *ordinal* requiere que los objetos de un conjunto puedan ser ordenados por rangos respecto a una característica o propiedad operacionalmente definida. El llamado postulado de transitividad debe cumplirse: si *a* es mayor que *b*, y *b* es mayor que *c*, entonces *a* es mayor que *c*. Es posible utilizar otros símbolos o palabras en sustitución de "mayor que", por ejemplo, "menor que", "precede a", "domina a", etcétera. La mayor parte de la medición en la investigación del comportamiento depende de este postulado. Debe ser posible efectuar las proposiciones ordinales o de orden de rango, como la que se acaba de utilizar. Es decir, suponga que se tienen tres objetos, *a*, *b* y *c*, donde *a* es mayor que *b*, y *b* es mayor que *c*. Si es posible decir, de manera justificada, que *a* es mayor que *c*, entonces se cumple la principal condición para la medición ordinal. Sin embargo, hay que tener cuidado. Quizá parezca que una relación cumpla el postulado de transitividad, aunque en realidad no sea así. Por ejemplo, ¿es posible decir siempre que *a* domina a *b*, y que *b* domina a *c*, y, por lo tanto, que *a* domina a *c*? Piense en esposo, esposa e hijo. Piense también en las relaciones "ama", "gusta", "es amistoso con", o "acepta". En tales casos, el investigador debe demostrar la transitividad. El procedimiento puede generalizarse de tres formas.

Primero, cualquier número de objetos de cualquier tipo puede medirse de forma ordinaria simplemente por medio de extensiones de *a*, *b*, *c*,..., *n*. (Aun cuando dos objetos algunas veces sean iguales, es posible realizar una medición ordinal.) Simplemente es necesario afirmar que $a > b > c > \ldots > n$, respecto de alguna propiedad.

La segunda extensión consiste en el uso de propiedades combinadas o criterios combinados. En lugar de usar solamente una propiedad, pueden usarse dos o más. Por ejemplo, en lugar de ordenar por rangos a un grupo de estudiantes universitarios respecto al rendimiento académico por el promedio de calificación, tal vez se desee ordenarlos por rangos respecto al criterio combinado de promedio de calificación y puntuaciones de prueba. (Los promedios de calificaciones también son puntuaciones compuestas.)

La tercera extensión se logra utilizando criterios distintos a "mayor que". "Menor que" es la primera que se piensa. "Precede a", "está por encima de", y "es superior a" son criterios útiles. De hecho, es posible sustituir símbolos por otros que no sean ">" o "<". Uno de ellos puede ser, "*O*", que puede utilizarse para significar cualquier operación, como las que se acaban de nombrar, donde se cumple el criterio de transitividad: *a O b* puede significar "*a* precede a *b*" o "*a* está subordinada a *b*", y *a O b O c* puede significar "*a* es superior a *b*, *b* es superior a *c*, y *a* es superior a *c*".

Los valores numéricos asignados a los objetos ordenados se llaman *valores de rango*. Sea *R* igual al conjunto de *objetos ordenados*: $R = \{a > b > \ldots > n\}$. Sea R^* igual al conjunto de *valores de rango*: $R^* = \{1, 2, \ldots, n\}$. Los objetos de R^* se asignan a los objetos de *R* de la siguiente manera: al objeto más grande se le asigna 1, al siguiente en tamaño 2, y así sucesivamente, hasta el objeto más pequeño, al cual se le asigna el último valor numérico de las series particulares. Si se utiliza este procedimiento, los valores del rango asignados aparecen en orden inverso. Si, por ejemplo, existen cinco objetos, donde *a* es el más grande, *b* el siguiente hasta *e*, el más pequeño, entonces:

Objetos	*R*	R^*
a	1	5
b	2	4
c	3	3
d	4	2
e	5	1

Por supuesto, es posible saltarse un paso si se asigna la R^* directamente: asignando 5 para *a*, 4 para *b*, hasta 1 para *e*.

Los números ordinales indican un orden de rango y nada más. Los números no indican cantidades absolutas ni indican que los intervalos entre los números sean iguales. Por ejemplo, no puede suponerse que, debido a que los *valores numéricos* estén igualmente espaciados, las propiedades subyacentes que representan estén también igualmente espaciadas. Si dos participantes tienen los rangos 8 y 5, y otros dos participantes los rangos 6 y 3, no es posible afirmar que las diferencias entre el primero y segundo pares sean iguales. Tampoco hay manera de saber que algún individuo no posea la propiedad que se está midiendo. Las escalas de orden de rango no son iguales a las escalas de intervalo ni tampoco tienen puntos con cero absoluto.

Medición de intervalo (escalas)

Las *escalas de intervalo* o *de intervalos iguales* poseen las características de las escalas nominales y ordinales, especialmente las del orden de rango. Además, las distancias numéricamente iguales en las escalas de intervalo representan distancias iguales en la propiedad que se mide. Así, suponga que se midieron cuatro objetos en una escala de intervalo y que se obtuvieron los valores 8, 6, 5 y 3. Entonces es posible afirmar legítimamente que las diferencias entre el primer y tercer objeto, respecto a la propiedad medida, $8 - 5 = 3$, es igual a la diferencia entre el segundo y cuarto objetos, $6 - 3 = 3$. Otra forma de expresar la misma idea de intervalo consiste en decir que los *intervalos* se suman y se restan. Una escala de intervalo se expresa de la siguiente manera:

a	*b*	*c*	*d*	*e*
1	2	3	4	5

El intervalo desde *a* hasta *c* es $3 - 1 = 2$. El intervalo desde *c* hasta *d* es $4 - 3 = 1$. Es posible sumar estos dos intervalos $(3 - 1) + (4 - 3) = 2 + 1 = 3$. Ahora note que el intervalo desde *a* hasta *d* es $4 - 1 = 3$, o expresado en una ecuación, $(d - a) = (c - a) + (d - c)$. Si estos intervalos fueran cinco alumnos medidos en una escala de intervalo de rendimiento, entonces las diferencias de rendimiento entre los alumnos *a* y *c*, y entre *b* y *d*, serían iguales. Sin embargo, no puede decirse que el rendimiento de *d* fue dos veces más grande que el del alumno *b*. (Dicha afirmación requeriría de un nivel más alto de medición.) Observe que no son *cantidades* o montos lo que se suma y resta, sino que son *intervalos* o *distancias*.

Uno de los ejemplos más conocidos de la escala de intervalo es la escala de temperatura centígrada o Celsius, que tiene un punto cero arbitrario donde el agua se congela, y un número 100 arbitrario en el que el agua hierve. Los puntos intermedios pueden dividirse igualmente utilizando la expansión del mercurio en un termómetro. Unidades iguales a lo largo de la escala representan cantidades iguales de expansión del mercurio. Como no se tiene un punto de cero absoluto, no se puede afirmar que 100° centígrados representen el doble de calor que 50° centígrados. Sin embargo, es posible afirmar que la diferencia o distancia entre 100° y 75° es igual a la diferencia entre 50° y 25°.

Tal como mencionó Comrey (1976) es mucho más difícil para las ciencias sociales y del comportamiento probar unidades de medición iguales que para las ciencias físicas o naturales. Los datos recolectados en las ciencias sociales y del comportamiento no están tan bien definidos como los datos de la temperatura. Lo que las ciencias sociales y del comportamiento intentan hacer es obtener mediciones que tengan una distribución normal (curva de campana). Si el instrumento de medición puede hacer esto, entonces se

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.

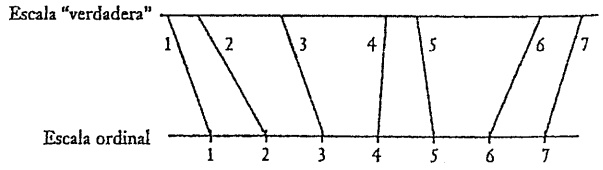
Comparación de escalas: consideraciones prácticas y estadísticos

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

que pueden utilizarse estadísticos como r, t y F con lo que, en efecto, son medidas ordinales? Y sí, se utilizan sin ningún reparo por la mayoría de los investigadores. Una de las excepciones es Cliff (1996), quien considera que los datos de las ciencias sociales y del comportamiento son, en el mejor de los casos, ordinales, y como tales únicamente deben utilizarse métodos ordinales de análisis de datos.

Aunque para algunos éste es un punto discutible, la situación no es tan difícil como parece. Como Torgerson (1958) señala, algunos tipos de origen natural han sido diseñados para ciertos tipos de medición. Para medir preferencias y aptitudes, por ejemplo, los puntos neutrales (a cada lado de los cuales hay grados positivos y negativos de favorecer, aprobar, gustar y preferir), pueden considerarse orígenes naturales. Además, las escalas de razón, aunque deseables, no son absolutamente necesarias, ya que la mayor parte de lo que se necesita hacer en la medición en psicología puede realizarse con escalas de intervalos iguales.

La falta de intervalos iguales es más seria debido a que las distancias *dentro* de una escala, en teoría, no pueden sumarse en ausencia de igualdad entre los intervalos. Aun así, aunque la mayoría de las escalas psicológicas son básicamente ordinales, puede suponerse, con considerable certeza, la equidad de los intervalos. La discusión es evidencial. Si se tienen, por ejemplo, dos o tres medidas de la misma variable y todas estas medidas son sustanciales y están relacionadas de forma lineal, entonces puede suponerse que hay intervalos iguales. Dicha suposición resulta válida, pues cuanto más lineal sea una relación, más cercanos a la igualdad serán los intervalos de las escalas. Esto también se aplica, al menos en algún grado, a ciertas medidas psicológicas como escalas y pruebas de inteligencia, rendimiento y actitud.

Un argumento relacionado es que muchos de los métodos de análisis utilizados trabajan bastante bien con la mayoría de las escalas psicológicas. Es decir, los resultados que se obtienen del uso de escalas y la suposición de intervalos iguales, son bastante satisfactorios. El punto de vista adoptado en este libro, entonces, es pragmático: que el supuesto de la igualdad de intervalos funciona. Aun así, se enfrenta un dilema: si se utilizan medidas ordinales como si fueran medidas de intervalo o de razón, quizá haya errores en la interpretación de los datos y en las relaciones inferidas a partir de los mismos, aunque el peligro probablemente no es tan grave como se ha hecho parecer. No hay problema con los números como tales, pues ellos no saben la diferencia entre p y r, o entre estadísticos paramétricos y no paramétricos; tampoco conocen los supuestos subyacentes. Pero nosotros sí sabemos, o debemos saber, las diferencias y las consecuencias de ignorar las diferencias. Por otro lado, si se acatan las reglas de forma estricta, se eliminan modos poderosos de medición y análisis, quedando sólo herramientas inadecuadas para enfrentar los problemas que se desean resolver (véase Nunnally, 1978; Comrey y Lee, 1995).

¿Cuál es la respuesta, es decir, la resolución del conflicto? Parte de la respuesta ya se mencionó: es probable que la mayoría de las escalas psicológicas y educativas se aproximen bastante a la igualdad de intervalos. En aquellas situaciones donde existan serias dudas sobre la igualdad de los intervalos, existen estrategias técnicas para enfrentar algunos de los problemas. El trabajador de investigación competente debe saber algo sobre métodos de escalación y ciertas transformaciones que cambian las escalas ordinales en escalas de intervalo (véase Barlett, 1947; Guilford, 1954, cap. 8; y Li, 1957). El tema de las transformaciones y sus propósitos y usos es importante; pero los científicos sociales y del comportamiento no le han dado la atención que merece. La mayoría de las referencias recientes respecto a las transformaciones provienen de estadísticos aplicados: Mosteller y Tukey (1977); Box, Hunter y Hunter (1978); Draper y Smith (1981); Box y Draper (1987), y Jennrich (1995). En lo que se refiere a las ciencias del comportamiento, las siguientes referencias cubren este tema: Cohen y Cohen (1983), Gorsuch (1983) y Howell (1997).

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).

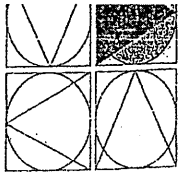
Resumen de capítulo

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

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.

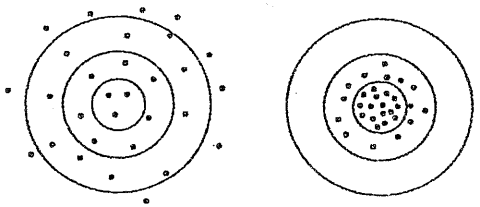
Lo mismo sucede con las mediciones en psicología y en educación: son más o menos variables de una ocasión a otra. O son estables o relativamente predecibles, o son inestables y relativamente impredecibles; son consistentes o no lo son. Si son confiables, entonces se puede depender de ellas; si no son 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



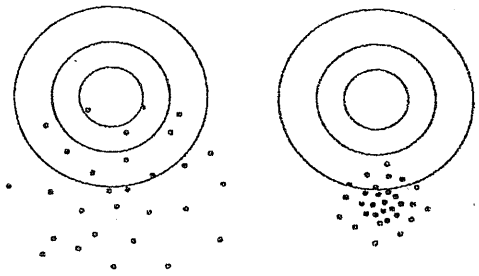
Rifle antiguo Rifle nuevo

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



Rifle antiguo Rifle nuevo

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 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.

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)

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:

$$X_t = X_\infty + X_e \qquad (27.1)$$

$$X_o = X_T + X_e$$

$$O = T + E$$

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:

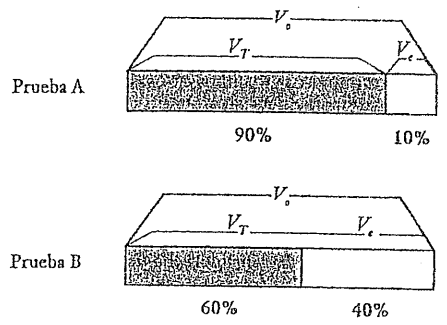
$$V_T = V_\infty + V_E \tag{27.2}$$

$$V_O = V_T + V_e$$

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_O, 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 $V_O = V_T$. 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:

$$r_n = \frac{V_\infty}{V_t} = \frac{V_T}{V_O} \tag{27.3}$$

$$r_n = 1 - \frac{V_e}{V_t} = 1 - \frac{V_e}{V_O} \tag{27.4}$$

donde r_n 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:

$$r_n = \frac{V_t - V_e}{V_t} = \frac{V_O - V_e}{V_O} \tag{27.5}$$

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_a). 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_e. Utilizando el análisis de varianza es posible calcular tanto V_t como V_e. 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 $r_{tt} = V_{ac} / V_{ind}$. Si se tuviera un camino directo para calcular V_{ac}, 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_{ind} para producir un estimado de V_{ac}. En efecto, es posible ignorar V_{ac} y restar la proporción V_e/V_{ind} de 1 y obtener r_{tt}. Ésta es una forma perfectamente aceptable para calcular r_{tt} y para conceptualizar la confiabilidad. La lógica de $V_{ind} - V_e$ 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_i	$\Sigma_{impares}$	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_a	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

Entre reactivos = $\frac{1\,190}{5} - 231.20 = 6.80$

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_i	$\Sigma_{impares}$	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_a	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

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_∞ deben ser iguales estadísticamente. Pero si V_o, 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_∞. 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:

$$r_{tt} = \frac{V_\infty}{V_{ind}} = \frac{V_{ind} - V_e}{V_{ind}} = 1 - \frac{V_e}{V_{ind}}$$

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_{tt} 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 ar... 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 mitades 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:

$$r_n = \frac{nr_{..}}{1 + (n-1)r_{..}}$$

Para la estrategia por mitades, n se establece igual a 2. La $r_{..}$ es la confiabilidad por mitades, y la r_n 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_b 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

$$r_{tt} = 1 - \frac{V_e}{V_t} = 1 - \frac{V_e}{V_O} = 1 - \frac{.80}{2.80} = 0.71 \qquad r_{t\infty} = 0.845$$

$$r_{tt} = \frac{V_\infty}{V_t} = \frac{2.00}{2.80} = 0.71 \qquad r_{tt} = r^2_{t\infty} = (.845)^2 = 0.71$$

$$VE_{med} = V_t(1 - r_{tt}) = 2.80\ (1 - 0.71) = 0.81$$

$$EE_{med} = DE_t\sqrt{1 - r_{tt}} = \sqrt{VE_{med}}\sqrt{0.81} = 0.90$$

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:

$$VE_{med} = V_t\,(1 - r_{tt}) \qquad (27.7)$$

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 Ítem* 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

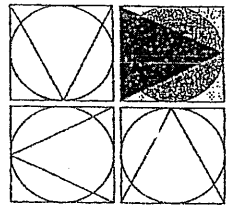
1. Este capítulo examina principalmente la teoría clásica de la confiabilidad. También contempla algunos de los desarrollos "más novedosos" en esta área.
2. La confiabilidad se define como la consistencia o estabilidad del instrumento de medición.
3. La teoría clásica de las pruebas creó la ecuación: $X_i = X_v + X_e$, donde X_i es la puntuación observada, X_v 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: $X_t = X_\infty + X_\varepsilon$?
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 será 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 (Referentes)	L_1	(.85)			
	C_1	–.07	(.88)		
Método 2 (Afirmaciones)	L_2	*.53*	–.15	(.81)	
	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 tradu cida 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

$$r_{tt} = \frac{V_{tv}}{V_t} \tag{28.1}$$

que es la proporción de la varianza "verdadera" entre la varianza total. Es teórica y empíricamente útil definir la validez de forma similar:

$$Val = \frac{V_{ta}}{V_t} \tag{28.2}$$

donde *Val* es la validez, V_{ta} 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.

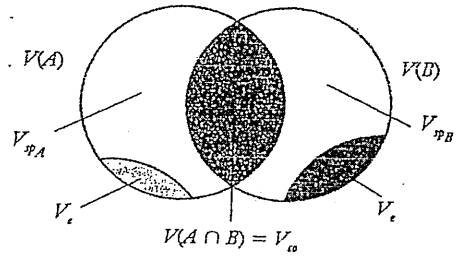
En contraste con la varianza del factor común de una medida está su *varianza específica*, V_{es}, la varianza sistemática de una medida que no es compartida por cualquier otra medida. Si una prueba mide habilidades que miden otras pruebas, entonces se tiene varianza de factor común; si también mide 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:

$$V_t = V_a + V_s + V_e \tag{28.3}$$

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



$$\frac{V_t}{V_t} = \frac{V_{co}}{V_t} + \frac{V_{es}}{V_t} + \frac{V_e}{V_t} \tag{28.4}$$

¿Cómo es que las ecuaciones 28.1 y 28.2 embonan aquí? El primer término a la derecha del signo de igual, V_{co}/V_t es el miembro derecho de la ecuación (28.2). Por lo tanto, l validez puede ser considerada como esa parte de la varianza total de una medida que no e varianza específica ni varianza del error, lo cual en forma algebraica se observa así:

$$\frac{V_{co}}{V_t} = \frac{V_t}{V_t} - \frac{V_{es}}{V_t} - \frac{V_e}{V_t} \tag{28.5}$$

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:

$$\frac{V_{co}}{V_t} = \frac{V_t}{V_t} - \frac{V_e}{V_t} - \frac{V_{sp}}{V_t} \tag{28.8}$$

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:

$$r_{tt} = \frac{V_t}{V_t} - \frac{V_e}{V_t} = \frac{V_{co}}{V_t} \tag{28.9}$$

Si se sustituye en (28.8), se obtiene:

$$\frac{V_{co}}{V_t} = \frac{V_{co}}{V_t} - \frac{V_{es}}{V_t} \tag{28.10}$$

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:

$$V_{co} = V_A + V_B \tag{28.11}$$

$$\frac{V_{co}}{V_t} = \frac{V_A}{V_t} + \frac{V_B}{V_t} \tag{28.12}$$

Entonces, utilizando (28.2) y sustituyendo en (28.12), se obtiene:

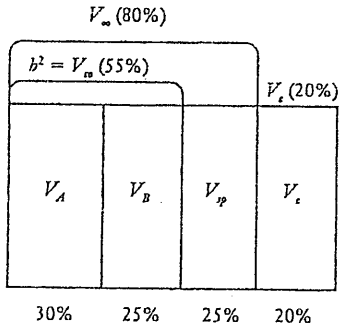
$$Val = \frac{V_A}{V_t} + \frac{V_B}{V_t} \tag{28.13}$$

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 demostra. se 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 $r_{tt} = (V_{ind} - V_r)/V_{ind}$ para I y II. Interprete las dos r_{tt}. ¿Por qué son tan diferentes?

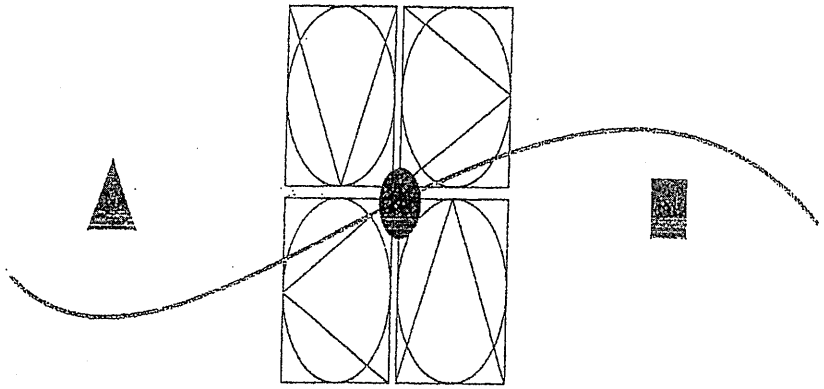
c) Sume los reactivos impares a través de los renglones; sume los reactivos pares. Compare los órdenes de rango y los rangos de los totales impares, de los totales pares y de los totales de los seis reactivos. Los coeficientes de correlación entre los reactivos impares y pares, corregidos, son .98 y .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

MEDIR

PARA

VIVIR

Markus Müller, Susana Ballesteros, Ma. Elena Bernal, Jaime Bonilla Barbosa, Joaquín Escalona, Ramón A. Gonzales, Katya Luna, Margarita I. Bernal-U.

Un pequeño pez lanza un chorro de agua que alcanza a una mosca en pleno vuelo. Un mono salta de una rama a otra rama sin caerse. Una libélula corrige constantemente la orientación de su movimiento para no perder de vista a su presa... La naturaleza mostraría así que cada especie tiene, en su forma de relacionarse con el entorno, una manera particular de hacer mediciones

MEDIR es algo que hacemos continuamente en la vida cotidiana. Cuando vemos, oímos, tocamos o cargamos un objeto, estamos realizando una medición. Las formas de medir en los seres vivos tienen una larga historia y han cambiado con la evolución. Estas formas alcanzaron su complejidad y exactitud máxima cuando la especie humana, HOMO SAPIENS, apareció en la tierra y desarrollo lo que ahora conocemos como la ciencia y tecnología.

Cuando al caminar por la calle vemos un charco, estimamos su tamaño para aumentar la longitud del paso y no mojarnos los pies. Medimos la distancia a al que se encuentra nuestro interlocutor y subimos o bajamos.

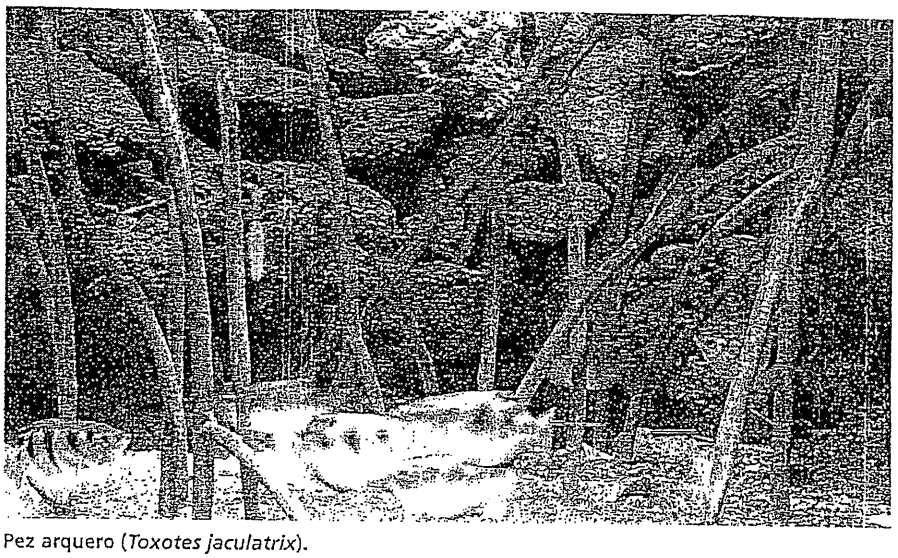
volumen de voz. Ajustamos la velocidad a la que cruzamos una calle de acuerdo con el tráfico. Para coordinar nuestros brazos y manos al tomar una bebida necesitamos estimar la distancia a la que están la taza o el vaso. El número de horas de sueño que necesita nuestro cuerpo depende de la edad de cada persona; es diferente en cada etapa de la vida. Este tiempo no depende de la latitud donde vivas; es decir, no depende del número de horas de iluminación solar, sino, al parecer, de un "reloj interno".

Nuestros sentidos son capaces de medir la composición de las sustancias específicas de un olor, así como de detectar la longitud y el peso de los objetos. Medimos la dirección —el ángulo— de donde viene un sonido y distinguimos casi instantáneamente entre el ruido, las palabras y una melodía de Mozart. Detectar las características cambiantes del entorno y evaluar los resultados de la percepción es sumamente importante para todos los seres vivos. Cada especie tiene, en su forma de relacionarse con el entorno, una manera particular de hacer mediciones.

El pez que caza desde el agua

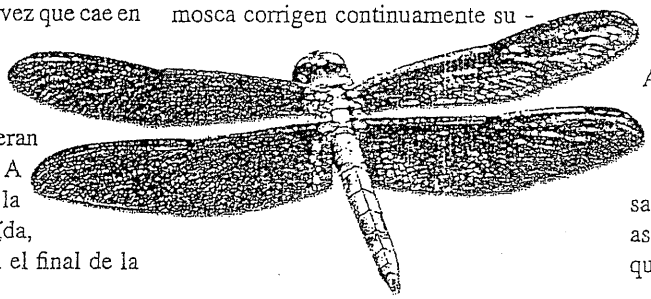
Un ejemplo extraordinario de lo anterior es el pez arquero, un pez pequeño de la familia *Toxotidae* que mide entre 10 y 20 centímetros de longitud y vive en los ríos y zonas de agua salobre del sureste de Asia. Este pez captura a sus presas desde la superficie del agua lanzando un chorro hasta una altura máxima de dos metros. Su boca funciona como una pequeña boquilla y su paladar tiene una ranura, que forma un tubo delgado cuando el pez le aplica la lengua. Cerrando las branquias rápidamente, el pez genera una presión a lo largo del tubito y el chorro de agua sale de la boca, como si ésta fuera una pistola de agua. Dispara su chorro a los insectos que buscan refugio para descansar bajo las hojas de las plantas que crecen pegadas al río. Como una mosca bajo la regadera, el insecto alcanzado cae arrastrado por el ímpetu del agua, trazando una trayectoria parabólica; y una vez que cae en el río, queda paralizado.

El pez tiene que ser rápido para cobrar la presa, pues sus atentos compañeros esperan atrapar también el bocadillo. A partir del momento en que la mosca da volteretas en su caída, el arquero se desplaza hacia el final de la trayectoria del insecto para recibirlo. El pez no tarda ni 100 milisegundos en desplazarse y lo hará sin echar siquiera un vistazo a la mosca durante el camino.



Pez arquero (*Toxotes jaculatrix*).

El pez se enfrenta a un problema típico de la física: el tiro parabólico. Para resolverlo sólo tiene que saber algunos datos: la posición y la velocidad inicial de la mosca en la caída. Un observador podría también conocer de manera precisa el punto donde la presa tocará la superficie del agua, haciendo un pequeño cálculo pero, por muy rápido que lo haga, el pez siempre lo encontrará antes. Si el pez fuera uno de nosotros, aparte de algunos conocimientos básicos de mecánica clásica, también tendría que saber algunos principios de óptica geométrica; sobre todo para asegurarse de dar en el blanco a la hora de lanzar el chorro. Hasta la fecha no se conoce ninguna otra especie con habilidades parecidas. Por ejemplo, un jugador de baloncesto también atrapa un objeto —el balón— que cae parabólicamente, pero no lo pierde de vista mientras éste va en el aire. Para atraparlo usa una técnica de retroalimentación: registra una y otra vez la posición del balón y va corrigiendo su dirección adecuadamente. Las libélulas atacan con una estrategia similar a la del basquetbolista. Durante la persecución de una mosca corrigen continuamente su orientación para asegurar que sean siempre las mismas celdas de sus ojos compuestos las que capten la imagen de su presa. De esta manera una libélula mejora continuamente la dirección de su vuelo, reduciendo cada vez más la distancia. Aunque la mosca vuele en una trayectoria muy sinuosa, la libélula mantiene el rumbo correcto.

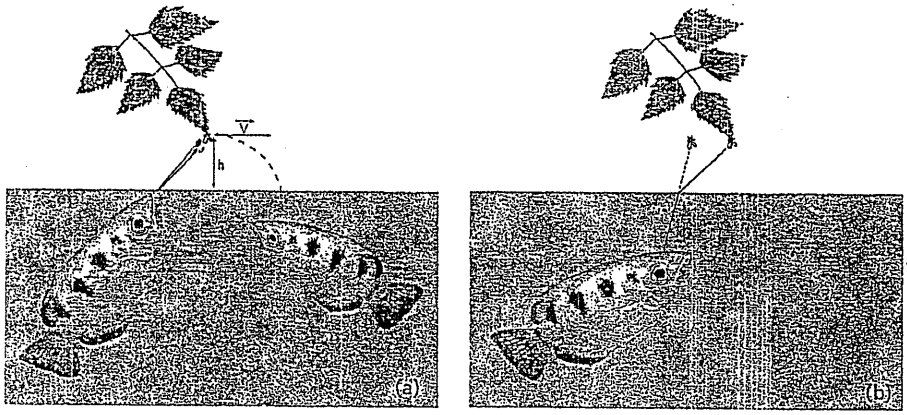


No cabe duda que el jugador de baloncesto y la libélula usan técnicas eficientes para lograr sus objetivos, pero la elegancia con la cual el pez arquero resuelve su problema hace sombra a los dos. No se sabe cómo puede este pequeño animal resolver un problema complejo de física clásica en una fracción de segundo. Aunque no creemos que el pez arquero conozca y mucho menos calcule la solución de las ecuaciones de la física clásica, este ejemplo sugiere que, al menos, tiene que realizar mediciones notablemente precisas para alimentarse.

¿Podemos entonces concluir que los mamíferos, los pájaros y quizá hasta los peces realizan mediciones? ¿Podemos decir que al registrar un conjunto de eventos y datos, además de experiencias, el sistema nervioso de los seres vivos está midiendo? ¿Es necesario ser consciente para realizar una medición y tomar una decisión a partir de los resultados?

Anzuelo de bacterias

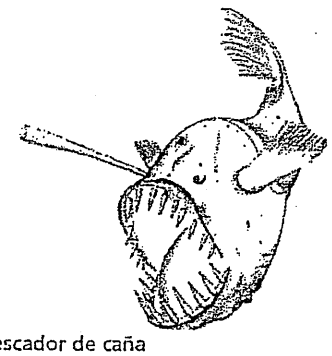
Aunque no podemos decir hasta qué grado tienen los mamíferos conciencia de sus actos (algo que a veces tampoco se sabe con las personas) ni hasta qué grado piensan o razonan, de las bacterias sí podemos asegurar, por su falta de sistema nervioso, que no piensan ni tienen conciencia. Sin



(a) El pez arquero tiene que calcular con precisión dónde caerá su bocadillo. (b) Debido al hecho que el agua tiene un índice de refracción mayor que el del aire, el pez ve a su presa en una dirección demasiado empinada, un pequeño detalle que tiene que tomar en cuenta para disparar en el ángulo correcto.

embargo, cuentan con un sistema de medición muy elaborado que, al igual que el de los organismos superiores, está desarrollado para asegurar que la especie sobreviva. Entre los ejemplos más sorprendentes están algunas bacterias que establecen una relación de cooperación (simbiosis) con varios tipos de animales marinos.

El pez pescador de caña vive en aguas marinas profundas y tiene un órgano especializado que se proyecta desde la parte superior de su cabeza y flota, a manera de anzuelo, sobre su gran boca. Su interior es un sitio ideal para el crecimiento de un tipo de bacterias, ya que ahí encuentran altas concentraciones de sus nutrientes preferidos. Las bacterias se sienten atraídas a entrar a este órgano y una vez ahí comienzan a reproducirse. Al cabo de unas horas, las bacterias alcanzan una densidad de población muy alta y empiezan a producir luz. El anzuelo luminoso atrae peces que se acercan sin darse cuenta a las enormes fauces del pez pescador y se convierten en su presa. El pez se beneficia de la presencia de la colonia de bacterias, al mismo tiempo que las bacterias aprovechan un medio rico en nutrientes. Pero, ¿cómo deciden



Pez pescador de caña

las bacterias que su densidad de población ha alcanzado el nivel adecuado para iniciar la producción de luz? ¿Necesitan hacer un censo de población por unidad de volumen?

Estas bacterias tienen una forma sorprendente de medir cuántos individuos de su misma especie habitan en su vecindario en un momento determinado. Aunque muchos dibujos sugieren que la superficie de una bacteria, es decir, la pared celular es lisa, en realidad es áspera, con una apariencia montañosa, llena de huequitos y pequeñas elevaciones. Algunos de estos huecos funcionan como enchufes para moléculas que se encuentran en su entorno y comúnmente se les conoce como receptores. Es decir, para cada tipo de molécula que tiene cierta importancia para las actividades de la bacteria en cuestión, hay "enchufes" específicos sobre su superficie que reconocen a esa molécula.

Una sustancia importante para estas bacterias, porque les sirve para comunicarse entre ellas, es un producto de su propio metabolismo. Una cierta concentración de bacterias produce una cierta concentración de esta sustancia. Estas moléculas se enchufan en los receptores y ocupan, en promedio, un número definido de receptores en la pared celular. Consecuentemente, el número medio de receptores ocupados es una medida directa de la concentración de la sustancia en un instante dado y, por lo tanto, de la concentración de bacterias.

Los receptores están conectados a unas estructuras interiores de la bacteria que controlan ciertas funciones de la célula; por ejemplo, la activación de genes particulares en su ADN. Cada vez que uno de los receptores está ocupado, se manda una señal a una de las estructuras interiores. Cuando la magnitud de la señal —es decir, la suma de todas las señales enviadas por el conjunto total de los receptores— llega a un valor umbral, la estructura interior activa una serie de genes encargados de producir los compuestos necesarios para generar luz. Así, las bacterias cuentan con un mecanismo refinado para medir el promedio de individuos por unidad de volumen. Estas bacterias cuantifican, es decir, miden un número (el promedio de receptores ocupados) que equivale directamente a la magnitud de su concentración, o densidad de población.

Unidades cambiantes

Así pues, no es necesario tener conciencia para hacer mediciones. ¿Y los seres humanos? ¿Somos siempre conscientes cuando percibimos información sobre nuestro entorno o tomamos decisiones? Cuando tocamos algo muy caliente retiramos rápidamente la mano y si súbitamente nos entra mucha luz en las pupilas, cerramos los ojos al instante. Esto se hace *sin* pensar,

Tiempo exacto

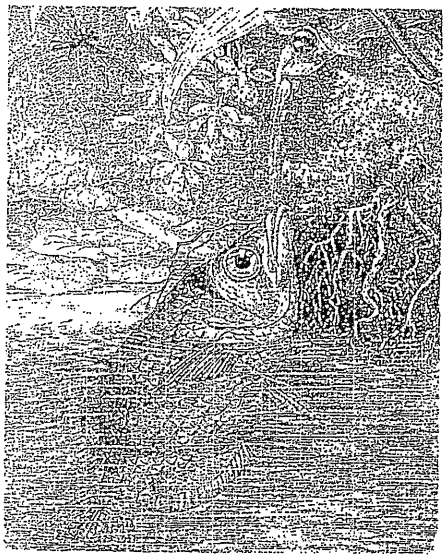
Un ejemplo del rigor que exigen los científicos al definir las unidades de medición es el desarrollo de la unidad básica de tiempo. Desde la antigua Babilonia se usa el día como unidad de tiempo. Pero esta definición basada en la periodicidad del día no es suficientemente precisa. Si se mide el lapso que dura un día en distintas épocas del año, se da uno cuenta que la unidad varía con las estaciones. Por si fuera poco, la rotación de la Tierra alrededor de su eje también tiene variaciones debidas a la fricción ocasionada por las mareas y a los desplazamientos del material de su interior. La fricción de las mareas (debidas a la interacción gravitacional entre la Tierra y la Luna) hace que la velocidad angular de la rotación de la Tierra alrededor de su eje disminuya continuamente y, por lo tanto, que la duración del día aumente.

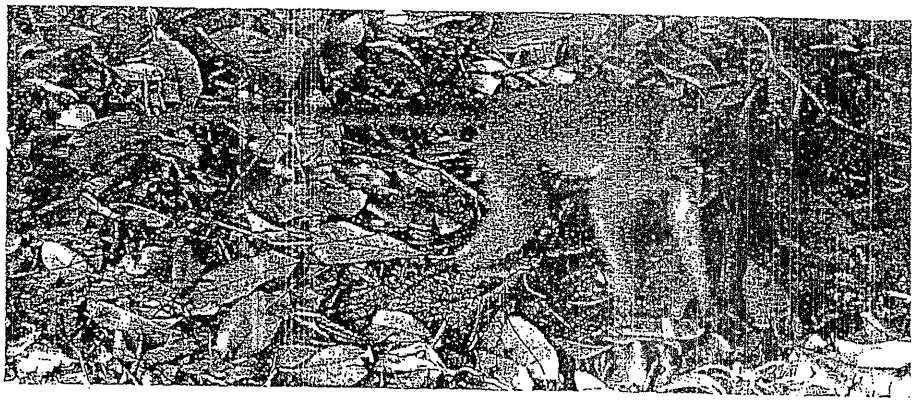
Para las aplicaciones científicas y tecnológicas es necesario establecer definiciones de unidades que no cambien con el tiempo y que sean objetivas; es decir, unidades que cualquiera pueda verificar obteniendo siempre el mismo resultado. Por esta razón la definición del segundo está basada en un sistema oscilatorio cuyo periodo es mucho más estable que la periodicidad de la rotación de la Tierra. Desde el año 1964 se usa la frecuencia de la luz emitida en ciertas condiciones por los átomos del Cesio 133. El segundo queda definido como la duración de 9192631770 oscilaciones de esta luz. La exactitud de los relojes de cesio más modernos es superior a una parte en 10^{-15} segundos. No se adelantan ni se atrasan más de un segundo en 20 millones de años.

o más bien s.. estar conscientes de lo que hacemos. Además, para asegurar que la respuesta sea suficientemente rápida, es necesario actuar *antes* de que el cerebro procese completamente la información y nos percatemos del riesgo. Sin embargo, no se puede negar que el reflejo, o la reacción del organismo, se debe a una medición (de la temperatura o de la intensidad de la luz).

Cuando se mide algo, sea una longitud, un ángulo o la concentración de un compuesto, se hace una comparación con algo conocido: una *unidad*. Dependiendo de los fines de la medición, esta unidad debe cumplir ciertos requisitos. Pero para la mayoría de las mediciones que realizan los seres vivos la definición de la unidad (su magnitud) no está bien establecida y basta tener una imagen intuitiva de ella. Por ejemplo, cuando un mono "mide" la distancia a una rama lejana, compara la distancia a la que se encuentra ésta con la distancia máxima que, según él, puede saltar. La magnitud de esta distancia de referencia está en su memoria. A través de la comparación, el mono decide con más precisión si puede alcanzar la rama de un salto o no. Se puede decir que el mono intuye en cierta forma si logrará realizar el salto. Las bacterias del ejemplo anterior usan una técnica de medición indirecta. En vez de contar el número de individuos por unidad de volumen, comparan la magnitud de la señal que llega a una estructura celular interior con un valor umbral. En estos procesos la comparación es crucial.

Hasta aquí podríamos concluir que medir significa detectar la magnitud de una





cantidad al compararla con algo conocido. La unidad de medida o de comparación puede ser un objeto, un recuerdo o un *umbral interno* del organismo.

Pero para tomar conciencia del resultado de una medición el tamaño de la unidad debe ser adecuado. Lo que percibimos fácilmente son conjuntos pequeños de objetos. Es decir, con sólo una mirada rápida distinguimos entre un conjunto de tres dulces y uno de cuatro. Si el número de dulces es 11 nos será más difícil notar si alguien nos robó uno. El problema se vuelve imposible de resolver de un solo vistazo, si el número de dulces es 23 o 27. Sabemos intuitivamente lo que significa un ramo de cinco, o quizá de 10 rosas rojas, pero, ¿podemos imaginar en forma precisa un ramo de 57 rosas rojas, o uno de 163 o de 132 854 945 rosas? La respuesta es no. Nuestro cerebro permite imaginar solamente números de objetos que son considerablemente pequeños. La exactitud de la percepción decae rápidamente cuando el número de objetos aumenta. Para evitar números muy grandes o fracciones muy pequeñas como resultado de la medición, se necesitan unidades de diferentes tamaños, como el angstrom (10^{-10} metros; con esta unidad se mide el tamaño de los átomos), el metro y el año luz (la distancia que recorre la luz en un año, unos 9.5 billones de kilómetros). Intuitivamente queda claro que la unidad metro no es la unidad conveniente si se desean medir distancias astronómicas, como la distancia entre dos galaxias, y tampoco sería muy útil medir las dimensiones de una molécula en metros. En general, el tamaño de la unidad debe ajustarse a la escala del objeto que se desea medir.

De estas consideraciones se concluye:

La comparación es la acción básica de la medición.

- La medición requiere una unidad de medida conocida para realizar la comparación.
- Para percibir el resultado de la medición, el tamaño de la unidad deber ser del mismo orden de magnitud que lo que se desea medir.

Estos son los requisitos necesarios para realizar una medición. No importa si se habla de una medición sumamente precisa en un laboratorio de investigación con alta tecnología o de la medición que realiza una bacteria; el momento de comparación en ambos casos es el elemento central.

Sin embargo, hay una diferencia importante entre las mediciones que realizan los seres vivos en la vida cotidiana y las mediciones con fines científicos. El tamaño de la presa que puede comer un pez depende naturalmente de su propio tamaño y la distancia que puede saltar un mono aumenta cuando éste crece. Así, además de que la magnitud de la unidad de medición en la mayoría de los ejemplos presentados no está bien definida, las unidades usadas por los seres vivos dependen del tiempo y continuamente deben ajustarse a las necesidades momentáneas del organismo. En cierta forma se puede decir que la unidad cambia con la evolución de cada especie y varía entre cada uno de los individuos.

Esta variabilidad temporal de las unidades, una propiedad que fue tan importante para la evolución de la vida en la Tierra, es justamente lo que se trata de eliminar para todos los fines científicos y tecnológicos.

Todos los autores de este artículo son académicos de la Universidad Autónoma del Estado de Morelos y responsables del proyecto del Museo de Ciencias en Morelos.

METODOLOGÍA OBSERVACIONAL

MTRA. MA. CELIA ESPINOSA ARÁMBURU
FACULTAD DE PSICOLOGÍA, UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO, 1997.

Contenido

Espinosa Arámburu, M. C. (1997). *Metodología Observacional.* México/ Facultad de Psicología UNAM.

Desarrollo histórico

El análisis experimental de la conducta es una ciencia que se establece a partir de principios y supuestos teórico-filosóficos derivados, en primera instancia, del positivismo lógico de la filosofía; en segundo lugar, del operacionalismo de la física y finalmente; del estudio de funciones de los mecanismos, de la biología (Catania, 1974). En cuanto a la metodología, ésta hubo de desarrollarse como consecuencia lógica de sus postulados científico-filosóficos, entre otras cosas porque la metodología prevaleciente de la llamada psicología experimental era de tipo introspectivo y presentaba problemas más que soluciones (Skinner, 1972).

Skinner (1972) desarrolla, a partir de sus predecesores como Hull, Watson, Thorndike y otros, una metodología de corte experimental, donde define como su objeto de estudio al comportamiento. Entonces, el interés primordial de la ciencia conductual se centra en el estudio del comportamiento del organismo y de sus interacciones con el medio ambiente (interno y/o externo). Esta ciencia se caracteriza por (Catania, 1974):

a. El empleo del método experimental, a fin de demostrar su validez para el tratamiento (experimental o aplicado) de la conducta.
b. La explicación, interpretación y predicción de las interacciones entre los acaecimientos medioambientales que controlan la conducta y esta última, como una contrapropuesta al tratamiento estadístico que describe la conducta de grupos de organismos.
c. El método, que se adapta al análisis del proceso conductual básico, es decir, el análisis de la conducta se concentra, en cada ocasión, sobre un número limitado de relaciones, acaecimientos ambientales y conductas, evitando la contaminación a causa de variables extrañas no controladas.
d. El uso de escenarios restringidos dentro de los cuales el organismo se comporta.

Ensayo elaborado para la materia de Análisis Experimental de la Conducta. División de Estudios Profesionales, Coordinación de Procesos Básicos y Metodología (Experimental), Fac. Psicología, UNAM. Enero de 1997. Revisión a cargo del Dr. Carlos Santoyo Velasco.

En este sentido, el análisis de la conducta centra su interés en el estudio del comportamiento del organismo como tal, por ende se estudia el comportamiento de un individuo, partiendo de que "lo importante no es estudiar a 100 individuos una hora, sino es estudiar a un individuo 100 horas" (Skinner, 1972). Este planteamiento difiere de la forma que ha seguido, tradicionalmente, la psicología experimental clásica; esta última dedicada al estudio del comportamiento de grupo o del hombre estándar o promedio (Skinner, 1970).

Por lo que, el analista de la conducta ha desarrollado una metodología experimental, sustancialmente, diferente a la empleada por el psicólogo experimental clásico.

Así pues, en la tradición de la Ciencia de la Conducta (Psicología Conductual), se han desarrollado diversa estrategias y herramientas metodológicas que permitan recabar información sobre el

organismo como individuo; indagando acerca de las relaciones existentes entre lo que hace el organismo y los factores medioambientales que determinan ese comportamiento (Skinner, 1979).

Hasta aquí se han planteado dos importantes argumentos que fundamentan la metodología empleada en esta ciencia. Primero, el estudio de centra en lo que hace un organismo como un todo; segundo, en como estas acciones están determinadas por factores ambientales a lo largo del tiempo. Tercero, la sustentación dada por la metodología experimental en esta área científica se consolidó en el laboratorio, donde, en principio, fue factible estudiar el proceso que recibió el nombre de condicionamiento reflejo; para después adecuarse al estudio de otro proceso al que Skinner (1979) denominó condicionamiento operante. En ambos casos, se requirió del desarrollo de instrumentos capaces de describir y registrar cualquiera de las relaciones existentes en ambos procesos y los posibles determinantes.

Para Skinner (1972) existen dos grandes clases de elementos representativos de esta relación, a saber la clase de estímulo como la unidad que representa al medio ambiente interno y/o externo, y la clase de respuesta que representa al comportamiento o ejecución del organismo como un todo. El modelo explicativo que se deriva de lo anterior descríbase las relaciones entre el medio ambiente y el organismo según la siguiente ecuación:

$$Y = f(X)$$

Donde Y es el comportamiento del organismo como un todo, f la función y X los determinante medioambientales. La ecuación se leería así: El comportamiento de un organismo (Y) está en función (f) de los determinantes medioambientales (X).

¿Por qué asumimos que el tipo de relaciones estudiadas dentro del Análisis Experimental de la Conducta son del tipo funcional? En principio, en la biología se ha estudiado a los organismos individualmente y se supone que el organismo es una máquina y que de ésta el interés central es su función y los mecanismos involucrados (Wartofsky, 1978). Al igual que la biología, la ciencia conductual lleva a cabo un análisis de carácter funcional (Skinner, 1970), dado que asume tal supuesto.

Resumiendo, el Análisis Experimental de la Conducta, elevado al estatus de la ciencia, busca estudiar las relaciones funcionales entre el medio ambiente y el organismo, a fin de encontrar y formular la ley natural en científica, y consolidar un sistema de leyes interrelacionadas que permita la construcción de una teoría sólida (Rosemblueth, 1971), que le de soporte.

El Método Científico

Los métodos como diría Bunge (1975) "son los medios arbitrarios para alcanzar ciertos fines". En este sentido el método científico es un sistema de procedimientos o técnicas diseñadas para indagar o tratar un conjunto de problemas (Bunge, 1975).

Al igual que la ciencia en general, en la ciencia conductual cada método es relevante para el particular estudio de un problema y las circunstancias donde emerge. Si hiciésemos una reducción de los pasos generales del método científico aplicado a esta ciencia fáctica o factual, a groso modo tendríamos:

1. La observación del evento y su relación.
2. La descripción de esta relación tal como ocurre y las circunstancias o situaciones en las que se presenta (Método inductivo).
3. La descripción de esta relación a través de una generalización (Hipótesis) bajo el esquema del método deductivo, si bien Sidman (1960) afirma que en Análisis Experimental de la Conducta lo relevante es saber escuchar al dato y no encasillarlo a las suposiciones (hipótesis deductivas) que el investigador tiene sobre él.
4. La medición de tales eventos, relaciones y circunstancias.
5. La contrastación experimental de la relación y sus circunstancias o la contrastación de la hipótesis deductiva o conjetura.
6. La descripción de los resultados obtenidos de esta experimentación.
7. La conclusión a la que nos llevan los datos obtenidos.
8. Una vez demostrada la relación, se evaluará el grado de generalidad, a fin de
9. Indagar la ley natural y formular la ley científica
10. Indagar si es factible encontrar un sistema organizado de leyes a fin de construir una teoría.

¿Qué diferencia plantea hacer una generalización empírica (método deductivo), o bien, observar y registrar una relación particular entre eventos (método inductivo) dentro del método científico?

Los métodos deductivo e inductivo

La exposición breve de las dos formas globales de razonamiento y quehacer que contempla el método científico, dará una idea sobre las decisiones metodológicas de la ciencia conductual al trabajar en su objeto de estudio. La primera forma describe al método deductivo, que se caracteriza por ser un sistema que parte de una generalización empírica que se transforma en una o varias conjeturas, comúnmente llamadas hipótesis deductivas, para después ser contrastadas a través de la experimentación (Bunge, 1975).

La segunda forma refiere al método inductivo que se caracteriza por partir de un dato empírico particular llevado a contrastación experimental, después de observar y registrar el desarrollo de una relación, se identifican los posibles factores que la determinan, y se analizan sistemáticamente, a fin de encontrar la generalidad y legalidad de la misma.

Según se ha referido anteriormente, en el Análisis Experimental de la Conducta, las estrategias metodológicas se fundamentan, esencialmente, en el método inductivo, dado que se centra en el estudio de lo que un organismo individual hace y como se relaciona con su medio ambiente.

En resumen, la ciencia conductual estudia datos empíricos particulares, a fin de concluir en lo general, es decir, hallar la ley natural que los rige (método inductivo).

De este hecho se deriva que la metodología conductual se ha caracterizado por:

a. Hacer una descripción objetiva de los acontecimientos tal como ellos ocurren (Skinner, 1970), y su relación, para después,
b. Separar y clasificar los elementos de la relación, y finalmente,
c. Registrarlos a través de un sistema diseñado ex profeso. Todo ello dirigido a descubrir orden, uniformidad y regularidad en las relaciones estudiadas; y demostrar que estos hechos tienen una relación válida con otros. (Skinner, 1970)

El método observacional y el experimental

Dentro del método científico se han generado diversas estrategias adecuadas al tipo de ciencia, sea: exacta, fáctica o social. Para la ciencia fáctica, uno de los primeros pasos dentro del método científico es la observación. Antes de los años 80's, sólo se consideraba a la observación como un elemento o paso inicial o un instrumento primordial en el método experimental para la consecución del experimento propiamente dicho. Sin embargo, tanto en las ciencias biológicas p. ej. en la Etología, como en el Análisis Experimental de la Conducta, el Enfoque del Desarrollo de la Interacción Social, y las Teorías del Aprendizaje Social y la Modificación de Conducta y Análisis Conductual Aplicado (Altmann, 1974; Anguera, 1983, 1991; Bakeman y Gottman, 1989; Blurton Jones, 1972; Cairns, 1979; Hinde, 1970; Santoyo, y López, 1990; Santoyo, 1994; Skinner, 1970; Smith y Connolly, 1980), ésta se constituyó en una metodología en su propio derecho.

Para el estudio del comportamiento de los organismos (incluyendo al organismo humano) en el escenario restringido o en su hábitat natural se han requerido desarrollar estrategias metodológicas, básicamente, de carácter observacional (Anguera, 1983). Esta metodología ha permitido el estudio de:

a. Los mecanismos implicados en tal comportamiento, es decir, cómo funcionan los sistemas conductuales;
b. Qué factores determinan comportamiento y su desarrollo y
c. Cómo se despliega el flujo conductual momento a momento in situ (Bakeman y Gottman, 1989) o en escenarios restringidos (Catania. 1974).

Además, este planteamiento puntualiza que se estudia al individuo en interacción con su medio ambiente y no al grupo en su conjunto. El supuesto de que el individuo es miembro y representante de la especie (Hinde, 1977), permite argumentar la importancia de recabar información sobre los procesos conductuales incluyendo al proceso social a través de la metodología observacional.

Para ello, y según la pregunta de investigación, cada estudioso del tema ha desarrollado un sistema de categorización conductual (Altmann, 1974; Anguera, 1983; Blurton Jones, 1976; Santoyo y Espinosa, 1989; Santoyo, Espinosa y Bachá, 1994; Smith y Connolly, 1980) que permite indagar sobre tales elementos implicados en el desarrollo, mantenimiento, y eliminación del comportamiento del organismo.

Los sistemas de observación etológica han sido pautas para el desarrollo de nuevos sistemas de observación conductual no sólo en la biología (Lyt on, 1980; Smith y Connolly, 1980), sino también dentro del ámbito de la Psicología como en el Anáiisis Experimental de la Conducta (Skinner, 1972) y la Psicología del Desarrollo Social Interaccional (Cairns, 1979; Patterson, 1979; Santoyo, Espinosa y Bachá, 1994) entre otros enfoques.

Por otra parte, el método experimental dentro de la Psicología Conductual (o Análisis Experimental de la Conducta) parte de la observación de tres cambios fundamentales: el primero es un cambio en el medio ambiente, al que por su función en la relación recibe el nombre de estímulo discriminativo o señal. El siguiente es un cambio en el organismo, que se traduce en comportamiento observable; y por último, un nuevo cambio en el medio ambiente siguiendo al comportamiento; al que se denomina consecuencia, dado que es efecto del mismo (Ribes, 1972).

Cabe aclarar que no todas las relaciones entre el organismo y su medio ambiente, necesariamente se ajustan a la descripción Skinneriana de la triple relación de contingencias, como es el caso de la evitación no discriminada descrita por Sidman (1953, citado en Catania, 1974).

Para el estudio de la triple relación de contingencias, Skinner diseñó una serie de procedimientos a los que genéricamente se les denomina procedimientos de reforzamiento y extinción y castigo. Los que han empleado para: a) el establecimiento o la adquisición, b) el mantenimiento y, c) la eliminación del comportamiento del organismo.

Además, dado que el foco de atención del análisis de la conducta es el estudio de lo que un organismo individual hace, se diseñaron estrategias o planes experimentales donde el organismo es su propio control (Castro, 1977; Kratochiwill, 1978). A estos planes se les conoce como diseños de series de tiempo de Nel. En estos diseños se registra uno o varios tipos de conductas y el tratamiento general incluye la observación y el registro momento a momento de lo que hace el organismo antes, durante y después del procedimiento experimental al que es sometido.

Es ingenuo pensar que el analista de la conducta hace generalizaciones de los hallazgos obtenidos de un solo organismo y que por ende, estos hallazgos poco aportan a la ciencia. Si bien, en la investigación básica, en analista de la conducta observa, registra y trata a un solo organismo; éste asume que hay regularidad, repetitividad y orden en las relaciones entre el medio ambiente y el organismo, y puede repetir, ya sea por réplica directa o sistemática (Sidman, 1960), las condiciones con diversos organismos individuales y encontrar la generalidad de esos hallazgos.

Retomando el punto medular de este trabajo, basta decir que la interrelación entre método observacional y método experimental en esta área del conocimiento es estrecha, dado que el segundo se basa en el primero y no podría existir sin él. Aunque la metodología observacional en sí misma sea la base para la metodología experimental, los hallazgos no alcanzan el soporte suficiente para dar explicación a las relaciones observadas. Dado que, para demostrar el estatus de una variable como el determinante de algún cambio conductual se requiere, necesariamente, de la manipulación de tal determinante, evaluando el grado en que el comportamiento varió por efecto del primero (Bunge, 1975).

A continuación se describe: qué es la observación, cuál es la importancia del dato conductual, Cómo se desarrollan las "taxonomías" o catálogos conductuales, varios de los sistemas de observación y recogida de datos conductuales, cómo se obtiene la fiabilidad del dato observacional y la compaginación de los métodos según el objeto de estudio.

La observación *in situ* y en escenarios restringidos

La observación es, quizás, para la ciencia fáctica, el primer elemento válido, dentro del método científico, sobre todo para aquellos enunciados empíricos "a posteriori" (Bunge, 1969, Wartofsky, 1978).

La observación como estrategia para recabar datos conductuales se ha caracterizado por desarrollarse a partir de:

a. El objeto de estudio, es decir, aquello que se va a observar;
b. Parejas de observadores entrenados para la recogida del dato conductual in situ o en escenarios restringidos;
c. Las circunstancias o situaciones donde se da la observación del objeto; y
d. Los medios de observación o instrumentos y medidas para recabar el dato conductual (Anguera, 1983; Backeman y Gottman, 1989; Bunge 1975).

Según Bunge (1975), los enunciados de observación quedarían descritos como "w observa a Y bajo x con la ayuda de z". En ese trabajo, justamente, se describe z que representa los sistemas observacionales y de registro.

Por otra parte, la observación constituida como una metodología, conlleva la creación de criterios que coadyuven la validez y fiabilidad de la información recabada. Por lo que, el proceso de observación debe permitir:

a. Que el objeto observado sea de carácter público. Condición mínima necesaria para sustentar que la ciencia es objetiva, y donde la regla afirma que los resultados obtenidos por observación pueden reproducirse por otros especialistas en condiciones análogas;
b. Clasificar los resultados, es decir, ubicarlos en clases de comportamiento, generando uno o varios sistemas o taxonomías de categorías conductuales;
c. Demostrar la fiabilidad de las propiedades o clase observadas de los hechos o las relaciones;
d. Describir objetiva y operativamente al hecho y/o a la relación entre hechos para llegar a,
e. La descripción y búsqueda de la función en las relaciones estudiadas (Anguera, 1983; Bakeman y Gottman, 1989; Bunge, 1969; Wartofsky, 1978).

La observación en las ciencias conductuales, también es referida como observación naturalista (Seitz, 1983). Como metodología, ésta abarca un gran espectro con métodos abiertos y métodos cerrados; o más precisamente denominados como: observación asistemática y observación sistemática, respectivamente. Y entre ambos métodos se encuentra la observación semi-sistemática que representaría un eslabón o tránsito entre ambos métodos.

La observación asistemática o de método abierto, se caracteriza por no tener un objeto de observación particular, p. ej. Cuando se hacen registros como: el anecdótico, la historia de casos, o un diario. Además, en este tipo de observación no hay un plan predeterminado de observación, o

el observador puede tener una idea difusa sobre lo que ha de observar (Bakeman y Gottman, 1989; Seitz, 1983).

Ejemplo de la observación semi-sistemática es la descripción de un espécimen o ejemplar, donde el observador registra lo que un organismo hace a lo largo de un periodo predeterminado de observación (Seitz, 1983).

En tanto que, la observación sistemática o de método cerrado, se basa en la selección de una estrategia sobre algún aspecto de la conducta a registrar, además este método es la base para el desarrollo de métodos de muestreo sistemático de observación, y se caracteriza por:

a. Ganar precisión en el registro del comportamiento (Seitz, 1983),
b. Permitir la cuantificación del comportamiento al sistematizar las estrategias de observación,
c. Definir de antemano varias modalidades de conducta o registrar a través del diseño de códigos conductuales;
d. El observador registra cada ocurrencia de comportamiento, anotando el código predefinido (estos registros, genéricamente, reciben el nombre de protocolos o catálogos de códigos (Backeman y Gottman, 1989);
e. Diseño de formas de medida a fin de establecer la fiabilidad de los instrumentos (Backeman y Gottman, 1989).

Además, este tipo de métodos se caracteriza por registrar y cuantificar los datos conductuales a través de la observación directa y no requerir de la inferencia. Cuando estos datos son tomados de puntajes obtenidos a través de observación indirecta como es el caso de los puntajes arrojados, p. ej. en las pruebas que puntean indirectamente la ejecución de un individuo; en vez de recolectar de manera directa los puntajes de tal ejecución (Cooper, Heron y Heward, 1987). Tal es el caso del uso de instrumentos sociocognocitivos de competencia social o cualquier otro tipo de prueba indirecta, donde como se mencionó antes, se hace necesario el uso de inferencia para traducir este dato (Cooper, Heron y Heward, 1987).

Esta es una de las razones más importantes que fundamentan el surgimiento y cada vez mayor uso de la metodología observacional en las ciencias conductuales (Lytton, 1980; Patterson, 1979; Skinner, 1970; Smith y Connolly, 1980).

Estos métodos se extenderán en el apartado sobre sistemas de observación y registro.

El dato conductual y su importancia científica

Para el analista conductual es de capital importancia validar la generalidad y confiabilidad de sus datos. A este respecto Sidman (1960) recapitula sobre las razones que llevan al investigador básico o aplicado al estudio de ese dato.

Así como, la ley natural existe independientemente de que la conozcamos y la formulemos, la importancia del dato es independiente de los propósitos del investigador (Sidman, 1960).

El que no tengamos la medida adecuada de ese dato, o no sustente nuestras suposiciones referente a él, o su frecuencia, duración u otra característica del mismo no se ajuste a nuestras expectativas; esto no debe conducir al investigador - sobre todo al analista conductual - a desecharlo, o peor aún a promediarlo.

Una ventaja del analista conductual respecto al psicólogo experimental - que estudia conducta de grupo, o conducta estándar o promedio -, es que el primero ha desarrollado, y continua desarrollando, estrategias que le conduzcan a:

1. Tener un dato lo más fidedigno posible del hecho que representa;
2. Elaborar la escala de medición que refleje fielmente su dimensión;
3. Respetar las diferencias individuales;
4. Generar sistemas de categorías que permitan una recogida confiable del mismo; y
5. Generar estrategias de tratamiento y evaluación más precisas que la generada por procedimientos estadísticos.

Esta ventaja es capitalizada por el desarrollo de diversos sistemas de observación, recogida y categorización de los datos, y la metodología observacional incrementa esta posibilidad.

Como base para su segmentación, el dato conductual se constituye en una unidad de conducta. En el análisis experimental, esta unidad de conducta se define de acuerdo a criterios operacionales y discretos acerca de lo que el organismo hace, p. ej. la respuesta de palanqueo, iniciaciones prosociales, etc.; en tanto que, en caso del análisis conductual aplicando lo criterios de definición varían según el problema a tratar, siendo éste el foco de interés, la conducta puede dividirse en continua o discreta, y genéricamente recibe el nombre de conducta blanco (Cooper, Heron y Heward, 1987; Kratochwill, 1978).

A continuación se presentan algunos criterios a tomar en cuenta en la construcción de taxonomías conductuales.

Taxonomías conductuales y su construcción

Una de las ventajas derivadas de la observación sistemática se refiere a que la sistematización del registro refleje, en términos observables, toda la información contenida en conductas o eventos; de manera que no se produzca perdida de información (Anguera, 1983).

En este sentido, en primer lugar, la observación directa y sistemática requiere de la construcción y uso de catálogos predeterminados de códigos de conducta.

En segundo lugar, de la especificación de la situación bajo la cual se tomará el registro.

En tercer lugar, de establecer las condiciones o características que definen al periodo de observación, y a las sesiones de observación (Anguera, 1983). Finalmente, en cuarto lugar de la validación y confiabilidad de estos sistemas de observación sistemática (Bakeman y Gottman, 1989).

En principio, en la observación sistemática se establecen registros anecdóticos para después dar origen a la creación de una taxonomía. Una taxonomía es un instrumento (originalmente ideado por los biólogos) que sustenta la clasificación de los elementos o eventos. En el caso del analista conductual, esta clasificación se ejerce tanto en la conducta como en el medio ambiente.

Esta taxonomía requiere de un alto nivel de abstracción, dado que la fragmentación del flujo conductual o medioambiental refiere a cada elemento como perteneciente a una clase, y por tanto, a tal o cual categoría. El evento a clasificar será colocado en dimensiones o características que quedan englobadas en conjuntos a fin de que lo definan.

Si bien, la mayoría de los investigadores conductuales toman parte de conducta no definida o categorizada ampliamente, como un código o clase global de comportamiento, p. ej. el caso de estar en la tarea (on task) o no estar en la tarea (off task); generalmente, el analista a diferencia de los etólogos, clasifica el comportamiento del organismo en unidades discretas de conducta, excluyentes y exhaustivas. Esto permite incluir los elementos de segmentación más precisos.

Skinner (1972) plantea que al clasificar los elementos tanto ambientales como comportamentales en clases de estímulos y clases de respuesta, permite la fragmentación, no arbitraria, de ambos tipos de flujo. Así mismo menciona que para hacer una fragmentación o partición de conducta se debe atender a la amplitud y complejidad de la misma. Ya que, habrá conductas globales (o molares) que contenga conductas más simples (o moleculares) y constituyan una conducta compleja.

Esta partición, sobre todo el flujo conductual, permite conformar las unidades de observación que después se constituirán en categorías conductuales y su definición se caracteriza por ser lo más objetiva y precisa posible (Ribes, 1972), eliminando adjetivos como malo, bueno, tenía la intención, etc.

En el ejemplo anterior, la definición y dimensiones de las categorías manejadas refieren a características globales, en exceso, es decir qué es y no es la tarea, como cuando se habla de un todo o nada.

La taxonomía conductual o sistema de categorías conductuales se constituye a través de:

a. La obtención de un registro anecdótico (Anguera, 1983; Bakeman y Gottman, 1989; Lytton, 1980);
b. La clasificación, de acuerdo a criterios específicos de segmentación y delimitación del flujo conductual;
c. La construcción de la definición de la categoría conductual en unidades de observación (Anguera, 1983);
d. La asignación de códigos a cada categoría conductual (Bakeman y Gottman, 1989).

Estas taxonomías o catálogos conductuales se constituyen, primeramente, a través de delinear ¿cuál es la pregunta del investigador? (Bakeman y Gottman, 1989), este planteamiento describe y deriva los lineamientos sobre el tipo de conducta focal o flujo conductual a categorizar. Por lo tanto, el planteamiento de la pregunta debe ser claro.

En segundo término, determinar, de acuerdo a ella, las conductas de interés, de tal forma que no se elaboren catálogos extensos, complicados y vagos (Bakeman y Gottman, 1989).

En tercer término, cada categoría conductual requiere de definirse operacionalmente, contener los elementos constitutivos que le den el estatus de clase de respuesta (Skinner, 1970). En este último rubro, se habla de que una conducta simple o molecular está definida por un número limitado de elementos en tanto que una conducta compleja o molar representa un tipo de interacción entre mayor número de elementos (Santoyo y Espinosa, 1987).

Las categorías conductuales se constituyen bajo ciertos criterios:

a. En principio, se definen operacionalmente;
b. La fragmentación del flujo conductual no debe ser arbitraria (Skinner, 1970), se requiere de una identificación de los elementos que componen una clase particular de conducta;
c. Se construye la definición de la categoría conductual;
d. Se asigna uno y solo un código por categoría conductual;
e. La definición de la categoría conductual se desarrolla con base en características físicas observables concretas y/o sociales objetivas - a fin de no requerir inferencia alguna se evitan los aspectos sociales en la definición - (Bakeman y Gottman, 1989);
f. El sistema de códigos se estructura a manera de que el observador tenga la facilidad para registrarlo, de ahí que se requiera que sean exactos; y
g. En lo posible, las categorías conductuales y sus códigos serán excluyentes y exhaustivos, es decir, que se asigne un código a una conducta particular que la diferencie de las otras (mutuamente excluyentes) y que exista un código para cada una de las conductas definidas (exhaustividad) (Bakeman y Gottman, 1989).

Las ventajas de tener categorías definidas exhaustivamente y mutuamente excluyentes se refiere a que el observador tendrá mayor facilidad en registrar la conducta, y evitar confundirla con otra, además se reduce el tiempo que se emplea para el registro a través de códigos.

Un sistema o una taxonomía conductual transita básicamente por tres instancias:

a. El establecimiento de las condiciones de observación,
b. La complejidad del sistema o taxonomía de categorías conductuales y
c. El desarrollo del sistema a partir de una observación informal (Bakeman y Gottman, 1989).

A continuación se describirán varios sistemas de registro donde se emplean los sistemas o las taxonomías conductuales.

Sistemas de observación y recogida de datos conductuales

Como se mencionó anteriormente, el analista conductual utiliza la metodología observacional como un medio idóneo para recoger el dato conductual en diversos escenarios. Esta metodología se caracteriza por tomar la medición directa del comportamiento o ejecución de un individuo en la situación restringida o in situ.

En la ciencia conductual, la recogida del dato conductual se realiza la medición directa del mismo, tal es el caso del registro de las respuestas o los puntajes de conducta. Por ejemplo estos son agrupados en tasa de respuesta, o en porcentaje de respuesta (Cooper, Heron, and Heward, 1987). Estas medidas directas no requieren de inferencia, dado que ellas representan características particulares del comportamiento medido.

Por otra parte, la medición indirecta ha sido extensamente usada como un apoyo adicional; sin embargo, los datos obtenidos a través de ella, para su interpretación requieren de la inferencia; los puntajes en alguna prueba sociométrica cognoscitiva, de actitudes, etc., son ejemplos de una medición indirecta del comportamiento (Cooper, Heron, and Heward, 1987). Este tipo de medida no se tratará en este trabajo.

Un analista conductual (u otra persona que aplica estrategias derivadas del análisis experimental de la conducta, dentro de ámbitos como la escuela) cuando trabaja en escenarios naturales o aplicados debe hacer una consideración importante, la que se refiere a que las unidades de conducta a observar deben medirse directamente y por ende traducirse a productos permanentes (Cooper, Heron, and Heward, 1987).

Productos permanentes: son productos que resultan de las conductas y se cotejan como ítems tangibles o efectos medioambientales duraderos. Estos datos se colectan a través de la medición y registro que toman lugar después de que la conducta ha ocurrido.

Ejemplo de productos permanentes son: resolver un examen escribiendo la respuesta sobre el papel, escribir palabras, colorear, completar rompecabezas, ensartar cuentas, ampliar bloques.

Otros ejemplos son: medidas del metabolismo (p. ej. el indicador de glucosa en la orina), alguna conducta audio grabada o video grabada, o el registro automático o registro por escrito y cualificado (Cooper, Heron, and Heward, 1987).

Estos productos permanentes tienen las siguientes ventajas.

a. Los productos permanentes son el resultado de algún tipo de instrucción,
b. En muchas instancias de conducta donde no es posible generar productos permanentes se utilizará equipo audiovisual a fin de que sea factible medirlas,
c. De esta manera, aún cuando los observadores no hayan tenido experiencia previa con este tipo de conducta y donde ocurre, pueden obtener acuerdo y confiabilidad en el registro al usar repetidamente el audiovisual,
d. Además estos productos pueden trasladarse a términos numéricos (Cooper, Heron, and Heward, 1987).

Para colectar los datos de los productos permanentes se emplean dos reglas: la primera se refiere a la medición de los productos permanentes como una opción cuando cada ocurrencia de la conducta blanco, naturalmente, es un producto y la segunda regla se refiere a la medición de la conducta blanco para generar un producto permanente, cuando esta conducta naturalmente produce esos productos; este es el caso de uso de audio o de video para conductas sociales como la conversación (Cooper, Heron, and Heward, 1987).

Entonces la metodología observacional se desarrolla a partir de medir directamente el dato conductual. Y se ha clasificado en tres formas: a. observación sistemática, b. semisistemática, y c. observación sistemática (Anguera, 1983). La tabla 1, muestra una forma de dividir a la observación.

La metodología observacional agrupa un conglomerado de sistemas, estrategias, y cuantificación del comportamiento. Primero, se presentan algunos métodos de observación más empleados: **abiertos o no sistematizados; semi-sistemáticos y cerrados o sistemáticos** (Anguera, 1983; Seitz, 1983).

OBSERVACIÓN	Asistemática o Abierta	a.- Historia de caso b.- Diario
	Semi-sistemática	a.- Descripción de espécimen o ejemplar.
	Sistemática o cerrada	a.- Medición directa= Puntajes directos de conducta
		b.- Medición indirecta= Pruebas: académicas, sociométricas no conductuales

Tabla 1. Forma de clasificar la observación.

Observación no sistematizada o muestreo *ab libitum*.- se caracteriza por proporcionar una descripción simple, llana y narrativa de carácter cualitativo e informal, donde se recogen todas las características de la conducta, evento, situación o escena que se observa (Anguera 1983; Smith y Connolly, 1980). Ejemplos de este tipo de observación son el registro anecdótico, historia de caso o diario (Seitz, 1983; Smith y Connolly, 1980).

En la observación no sistematizada existen dos tipos de descripción que se encuentran, inadvertidamente, mezclados: la descripción molecular y la descripción molar. La primera registra conductas o eventos puntuales concretos en exceso, tal es el caso de las conductas motoras. La segunda, emplea un nivel de abstracción mayor, donde se hace necesaria la interpretación de un conjunto amplio de situaciones, eventos o conductas (Anguera, 1983).

Ejemplo de observación no sistematizada

9:13 Ana entra al salón, y la profesora le llama la atención; Ana responde que fue al baño (9:15). 9:16 La profesora le da instrucciones para que se siente y haga la tarea de colorear un dibujo de un elefante. 9:18 Ana se sienta y pregunta a la profesora ¿de qué color? (9:20). 9:24 Ana colorea el elefante.

Observación semi-sistematizada o descripción de espécimen.- Se define como aquel tipo de descripción donde se registra a un individuo ejemplar o espécimen a través del empleo de una

cámara; y donde el investigador puede tener una aproximación más precisa de lo que ha de observar y registrar al usar tecnología más sofisticada. A este tipo de observación se le denomina semi-sistematizada debido justamente a que carece de: a. un objeto de estudio definido; b. la selección (o desarrollo) de un sistema de observación, codificación y registro que proporcione los datos observacionales que describan el problema a estudiar (Anguera, 1983; Seitz, 1983).

Ejemplo de observación semi-sistematizada (tomado de Anguera, 1983):

Hora	Evento antecedente	Conducta central por observar	Evento consecuente
9:13		1. Ana entra al salón	2. Profesora llama atención
9:15		3. Ana responde	
9:16	4. Profesora de instrucción		
9: 18		5. Ana se sienta	
9:20		6. Ana pregunta	
9:22			7. Profesora responde
9:24		8. Ana colorea	

Métodos de observación sistemática

La observación sistemática se define como una vía específica accesible para registrar y cuantificar la conducta (Bakeman y Gottman, 1989). Los métodos de observación sistemática que presentaremos a continuación se caracterizan por permitir:

a. La observación y registro de la conducta, cuando no es conveniente o viable el uso de equipo especial,
b. La observación y registro de la ocurrencia espontánea de la conducta en contextos naturales,
c. La definición de antemano de varias modalidades de conducta y se registran los códigos asignados a ellas,
d. Un amplio grado de concordancia en la toma el dato conductual (Cooper, Heron, and Heward, 1987; Bakeman y Gottman, 1989).

Una consideración importante respecto a la observación sistemática es que ésta no siempre permite observar secuencias de conducta (Bakeman y Gottman, 1989) por lo cual el investigador habrá de ponderar, según su pregunta, que tipo de unidad conductual requiere observar, ya que si la pregunta se refiere a patrones o estados y sus diferencias tendrá que contemplar el uso de un sistema que permita la recogida de los datos secuenciales.

En esta sección se describirán algunos de los sistemas y procedimientos de medición directa, y registro de las conductas, desarrollados en la metodología observacional. Bakeman y Gottman (1989) mencionan que antes de seleccionar un sistema de registro observacional, el investigador necesita decidir el tipo de unidad o unidades de conducta que utilizará. Y para su codificación existen varias posibilidades: métodos de muestreos, métodos de observación y registro de eventos

o intervalos. La selección depende del número de factores acerca del tipo y la complejidad del sistema de codificación, se precisión y las características del equipo de registro disponible.

Cuando el analista de la conducta se refiere sólo a la frecuencia de la ocurrencia de ciertos eventos y su orden de presentación, se habla de la selección de unidades de registro por eventos momentáneos (o frecuencias de conductas), en tanto que, cunado el analista selecciona como medida la cantidad o proporción de tiempo ocupado o gastado por una conducta particular se registran "estados conductuales", o duraciones de conducta (Bakeman y Gottman, 1989).

Además, estas estrategias también hacen diferencia entre si el registro es continuo o intermitente (Bakeman y Gottman, 1989). En el análisis experimental de la conducta, dentro de los escenarios restringidos, en principio, se ha empleado el registro continuo, p. ej. el registro acumulativo inventado por Skinner (1970). Sin embargo, en escenarios naturales aparatos para el registro acumulativo, son poco usados y se prefiere usar papel, lápiz y cronómetro. Una característica importante de este tipo de registro es que permite realizar un análisis secuencial del flujo conductual (Bakeman y Gottamn, 1989). A continuación revisaremos brevemente los registros: de muestreo, de eventos, de duraciones, de latencias, de intervalo, y de muestreo de tiempo, entre otros.

Muestreo sobre una persona focal.- en este método se detalla completamente todas las conductas específicas observadas de cada sujeto. En algunas ocasiones se utiliza una medición de matriz sociométrica completa a fin de incluir observaciones adicionales en ciertos individuos (Altmann, 1974; citada por Smith y Connolly, 1980).

Registro de evento

Registro de evento.- Es el registro de cierta conducta cada vez que ocurre. El observador anota el número de veces que ocurre una conducta (blanco), usualmente el observador tomó la información adicional a través de la descripción (Cooper, Heron, and Heward, 1987; Bakeman y Gottman, 1989; Seitz, 1983). Para recoger el dato conductual en este sistema de registro se emplean y diseñan listas de control (checklist; Bakeman y Gottman, 1989).

Dentro de este tipo de registro se incluyen tres tipos de muestreo: a. de todo-nada o uno-cero, b. muestreo de todas las ocurrencias y, c. muestreo de escudriñamiento instantáneo.

El muestreo de uno-cero o todo-nada.- Dentro de un periodo muestra se registra la ocurrencia o no ocurrencia de las conductas observadas de un individuo particular (Altmann, 1974; citada por Smith y Connolly, 1980).

Inmerso en este tipo de muestreo se encuentra el muestreo de secuencias que se refiere a la observación de secuencias conductuales, p. ej. secuencias de interacción, de las cuales se registra su inicio y termino (Altmann, 1974; citada por Smith y Connolly; 1980).

Muestreo de todas las ocurrencias o muestreo incidental.- En este caso se observa a todo un grupo y se registra la ocurrencia de alguna conducta particular (Altmann, 1974; citada por Smith y Connolly, 1980). Este tipo de registro se usa más dentro de la Etología y el Análisis Conductual aplicado ya que se refiere a la observación de una conducta específica de los miembros de un grupo.

Muestreo de escudriñamiento instantáneo.- Se refiere al escudriñamiento de todo un grupo en intervalos regulares, la conducta particular de cada individuo es muestreada en breves periodos de tiempo (Altmann, 1974; citada por Smith y Connolly, 1980). Este sistema se usa sobre todo cuando el interés del analista conductual está centrado en obtener datos sobre una red grupal, dado que se inicia el registro con un sujeto particular, p. ej. por 5 segundos. A los siguientes 5 segundos de observación se registra la conducta del siguiente sujeto en la lista. El periodo de observación concluye cuando se tienen los datos de todos los miembros del grupo. A fin de describir y representar la red grupal se agrupan los datos obtenidos en sesiones o periodos repetidos a lo largo del estudio (Smith y Connolly, 1980).

Las ventajas de este sistema de registro son: a. no interfiere con otras actividades en marcha, b. el registro es fácil, c. produce productos permanentes. Este sistema presenta varias restricciones. La primera restricción se refiere a tratar sólo con unidades discretas de conducta definidas en su totalidad (descripción de principio a fin) p. ej. en el susurro que es difícil acordar donde inicia, o en el un tarareo cuando se inicia y finaliza uno y otro inicia.

La segunda restricción se refiere a que hay conductas que no ocurren en altas tasas, haciéndose difícil para un observador registrar adecuadamente cada ocurrencia discreta. O bien, cuando las conductas son difíciles de registrar debido a que se presentan a altas tasas, o cuando las conductas ocupan un amplio periodo de tiempo, p. ej. escuchar, jugar solo, etc.

El registro por evento, registra el número de ocurrencias de la conducta, sólo si existe la oportunidad de responder varias veces a lo largo de las sesiones de observación, y si este responder se mantiene constante (p.ej. cada sesión de observación de 10 minutos; Cooper, Heron, and Heward, 1987).

Ejemplo de una lista de control del registro de eventos (Adaptado de Bakeman y Gottman, 1989):

Observador_____	Sujeto_____
Tiempo de inicio_____	Fecha_____
	Tiempo final_____

Golpes	Peleas	Solicitud de ayuda
𝍸 𝍸	𝍸 𝍸	𝍸

Otro tipo de registro de eventos es el registro de secuencias de eventos, en él se registra cada cambio de estado conductual a lo largo del tiempo, mostrando una continuidad entre unidades de diferentes conductas, sucesivamente codificadas (Bakeman y Gottman, 1989). Este registro contiene parte del registro de evento y del registro por intervalo.

Ejemplo de un registro de una secuencia de eventos (Bakeman y Gottman, 1989):

Observador_____ Sujeto_____
Tiempo de inicio_____ Fecha_____
Tiempo final_____

Conducta 1: S=solo Conducta 2: P=paralelo Conducta 3: G= grupo

SPSPGSGPGPS...

El analista conductual debe ser capaz de determinar cuál es el método de registro más apropiado para recolectar el dato conductual.

Registro de duración.- Consiste en registrar la cantidad de tiempo total que el individuo emplea en realizar una amplia conducta (Cooper, Heron, and Heward, 1987). El registro de duración se emplea cuando el interés del analista de la conducta se centra en la amplitud en el tiempo ocupado por cada conducta (evento conductual o tasas de conductas continuas). Este registro es usado cuando, la conducta es emitida a altas tasas y estas altas tasas se agrupan en una más general que las incluye. Por ejemplo, abofetear la cara de alguien, cuando ocurre a latas tasas puede ser difícil y poco confiable que el observador identifique cuando inicia y termina cada unidad molecular de conducta; el registro de duración es el indicado. Si el abofetear ocurre a tasas moderadas y es factible identificar el inicio y el fin de la unidad conductual, se usará el registro de eventos general.

Este método de registro presenta dos procedimientos de recolección de datos conductuales: a. duración total o, b. duración por ocurrencia.

En el procedimiento de la duración total se registra durante el periodo de observación la cantidad de tiempo que un individuo ocupa en una conducta. Un criterio de selección del método de observación y registro de eventos o de duración total se basa en las diferencias existentes en las

dimensiones de la conducta. Otro criterio es la medición que se hace de las dimensiones conductuales sea esta la frecuencia o la dimensión temporal (Cooper, Heron, and Heward, 1987).

En el registro de la duración total, un cronómetro se activa cuando inicia o termina la conducta sin regresar el cronómetro. Es decir, se registra de manera repetida la ocurrencia de una conducta; anotando el tiempo ocupado por esa conducta sin detener el cronómetro hasta que concluye el periodo de observación.

El observador inicia el conteo del siguiente episodio conductual, de esta forma, el observador obtiene la duración acumulada a lo largo de la duración del periodo de observación que puede contener varios episodios conductuales. El tiempo ocupado por cada episodio conductual queda anotado en una hoja de registro.

El registro de la duración total frecuentemente se reporta de dos maneras. Un método es reportar la duración acumulada de la ocurrencia de la conducta a lo largo de un periodo específico de tiempo (Cooper, Heron, and Heward, 1987). A continuación se presenta la forma en que se obtiene el porcentaje.

$$\frac{Duracion\ total\ de\ conducta\ Y}{Duracion\ total\ del\ periodo\ conductual\ Z} \times 100 =$$

Donde el periodo de observación total de la conducta Z es el divisor de la operación, y el dividendo es el tiempo parcial ocupado por la conducta Y (particular), y el cociente es multiplicado por 100. Cuando el criterio de tiempo no específica un mínimo o un máximo se utilizará el registro de duración total, colocando la cantidad total individual del tiempo requerido para completar una tarea específica.

El método de registro de la duración por ocurrencia o registro de tiempos de inicio y fin. Se refiere a aquellos casos donde se registra, con un cronómetro, el tiempo de inicio y el término de la conducta.

La estrategia, aquí, consiste en activar el cronómetro cuando inicia el episodio conductual y desactivarlo cuan éste concluye. El observador transfiere la duración del tiempo mostrado en el cronómetro a una hoja de datos y se regresa el cronómetro. Nuevamente se echa a andar el cronómetro al iniciar la segunda ocurrencia de la conducta y se detiene al final del episodio, la duración del tiempo es transferida a la hoja de datos y el procedimiento se continúa hasta el final de la sesión de observación (Cooper, Heron, and Heward, 1987; Bakeman y Gottman; 1989).

Un criterio importante para el uso de este registro es el uso de categorías exhaustivas y excluyentes.

Ejemplo de la hoja de datos para el registro de la duración por ocurrencia (adaptada de Cooper, heron, and Heward, 1987):

Hoja de datos de la duración por ocurrencia

Nombre del sujeto:_____ Observador:_____

Conducta:_____

Fecha:_____

Tiempo de inicio:_____ Tiempo de término:_____

Condición:_____ Número de sesión:_____

Número de episodios	Tiempo transcurrido por episodio (en minutos ['] y segundos ["]
1	1'17"
2	6' 5"
3	2' 1"
4	3'35"
	Total 12'58"

Registro de latencia

Registro de latencia.- Consiste en la medición del tiempo que pasa entre el comienzo de un estímulo (p. ej. dirección de la tarea, señal) y la iniciación de una conducta.

El registro de la latencia puede usarse cuando el principal interés es la longitud del tiempo entre una oportunidad para emitir una conducta e iniciar la ejecución de esta conducta (Cooper, Heron, and Heward, 1987). Este registro sirve también para latencias cortas entre el inicio del estímulo y el inicio de la respuesta.

El procedimiento del registro de latencias es similar al registro de duración, en ambos:

a. Se colecta el dato acerca de la dimensión temporal de la conducta,
b. Se usa el mismo procedimiento de medición,
c. Se requiere la identificación precisa del fenómeno a registrar.

Ejemplo de la hoja de datos usada para el registro de latencia (adaptada de Cooper, Heron, and Heward, 1987):

Hoja de datos del registro de la latencia

Nombre del sujeto:_____ Observador:_____
Conducta:_____
Fecha:_____
Tiempo de inicio:_____ Tiempo de término:_____
Condición:_____ Número de sesión:_____

Número de estímulos	Tiempo transcurrido por episodio (en minutos ['] y segundos ["]
1	2'20"
2	3'58"
3	1' 2"

Registro por intervalo.- En este registro se hace la partición del tiempo de observación en unidades iguales o de longitud variable, entre sí. Aquí se especifica la clase de conducta que ha de ser observada durante un periodo de tiempo preseleccionado (Anguera, 1983; Bakeman y Gottman, 1989; Seitz, 1983). Po ejemplo, si el total del periodo de observación es de 10 minutos y el observador usa intervalos de 10 segundos cada uno, la sesión se dividió en 60 unidades iguales (Cooper, Heron, and Heward, 1987).

El observador diseña de antemano su hoja de registro dibujando una serie de cuadros o celdillas en papel. Cada celdilla representa un intervalo. Las celdillas pueden ser colocadas horizontal o verticalmente, como se muestra en el ejemplo.

El observador marca en cada intervalo de tiempo con un símbolo que indica si la conducta ha ocurrido. La colección de datos vía intervalo por tiempo es reportada como porcentajes de intervalos, frecuencias absolutas, frecuencias relativas o, la duración de la(s) conducta(s) observada(s).

Una ventaja que se deriva de la selección unidades del intervalo es que el observador pueda tener tiempo suficiente para observar y registrar confiablemente la conducta. El tamaño del intervalo usualmente va de un rango de 6 a 15 segundos (Anguera, 1983; Cooper, Heron, and Heward,

1987). El registro por intervalo es muy útil debido que sirve tanto para registrar altas o bajas tasas de conducta así como la duración de las mismas.

Una recomendación adicional para cualquier tipo de registro y en particular para el de intervalo se refiere a elaborar y colocar al inicio de la hoja de registro, la ficha de identificación. A manera de hacer más fácil el trabajo de recolección de los datos observacionales al describir a que sujeto se observó, la fecha en que se tomó el registro, el nombre del observador, el tiempo de inicio y termino del registro. Ingresando todos los elementos que le permitan al investigador hacer el cotejo de los datos observacionales, es decir, de las conductas asentadas en él.

Comúnmente, el registro por intervalos es de papel y lápiz, una tabla con sujetapapeles, y un cronómetro. El cronómetro puede estar prendido a la tabla. Una desventaja potencial del uso de tablas con sujetapapeles y cronómetro, es que el observador periódicamente debe mirar al sujeto y al cronómetro. Por lo que, quizás, decremente la oportunidad de medir la ocurrencia de la conducta.

Worthy (1968, citado por Cooper, Heron, and Heward, 1987) describe un aparato miniatura que genera señales audibles que puede ser usado para el registro de intervalos. Este aparato elimina la necesidad de monitorear el cronómetro.

Sin embargo, esta desventaja potencial queda resuelta cuando se cuenta con observadores entrenados y se exige una alta concordancia entre sus registro observacionales (Santoyo y Espinosa, 1987, 1994). Usar un instrumento o aparato que genere señales audibles como algún emisor de tonos o por cintas grabadas y reproducidas en las grabadoras audio-cassette con un audífono para el oído a mostrado bajar la concordancia entre observadores por efecto intrusivo de este tipo de aparatos en la investigación in situ (Espinosa y Santoyo, 1995).

Ejemplos de dos tipos de hoja del registro por intervalos (adaptado de Cooper, Heron, and Heward, 1987):

Hoja de datos del registro por intervalos

Nombre del sujeto:_____ Observador_____
Conducta_____
Fecha_____
Tiempo de inicio_____ Tiempo de término:_____
Condición:_____ Número de sesión:_____

Intervalos de 10
Presentación vertical

1	/	7	X
2	X	8	X
3	/	9	/
4	/	10	/
5	X	11	X
6	/	12	X

Intervalos de 10 segundos
Presentación horizontal

1	2	3	4	5	6
/	X	/	/	X	/
7	8	9	10	11	12
X	X	/	/	X	X

/ = El estudiante está en la tarea (on-task)
X = El estudiante no está en la tarea (off-task)

Los datos de la hoja de registro por intervalos muestra que el estudiante ha estado en la tarea durante el 50% de los intervalos.

$$\frac{(intervalos\ en\ la\ tarea\ [on-task])\ 6}{(total\ de\ intervalos)12} \times 100 = 50$$

El registro por intervalos se divide en: el registro de intervalo parcial y el registro de intervalo completo.

El registro de intervalo parcial.- El procedimiento inicial más común requiere que el observador registre simplemente si la conducta está o no presente en algún momento durante el intervalo. El

analista que usa este método no está interesado en cuanto tiempo dura la conducta en el intervalo o cual es su longitud.

Cuando el analista está interesado en conocer si la clase de conducta ha ocurrido por largo tiempo usará el registro de intervalo completo.

El registro de intervalo completo.- se emplea cuando la conducta se presenta en el intervalo completo, sólo si aparece y ocupa todo el intervalo se considerara que ésta ha ocurrido.

Muestreo de tiempo momentáneo.- Este método es similar al registro de intervalos. La diferencia estriba en que el muestreo de tiempo momentáneo se registra la presencia o ausencia de la conducta que sigue inmediatamente a un intervalo de tiempo específico. En el registro de muestreo momentáneo, el observador registra la conducta que toma lugar al final del intervalo. En este caso cada intervalo de tiempo es variable y se programa a través de la selección de números aleatorios según la duración del periodo o sesión de observación, estos intervalos se distribuyen a lo largo de una línea numérica, con el criterio de que ninguno de dos números sucesivos tuviese menos que 10 segundos. La selección de los números (tiempo de intervalo) se programa de manera similar al programa de reforzamiento de intervalo variable (Skinner, 1970). La conducta se registra una sola vez al final de cada intervalo.

> Por ejemplo un conductor de un programa recreativo de verano para niños de edad elemental estaba interesado en el juego cooperativo de un niño, durante el periodo de actividad de 60 minutos. El periodo fue dividido en intervalos de 5 minutos, esto permitió al conductor observar y registrar el juego cooperativo de los niños por 12 ocasiones. El conductor registró sólo la conducta en el instante en que finalizaba el intervalo. (Ejemplo citado por Cooper, Heron, and Heward, 1987).

En el siguiente ejemplo del registro por intervalos muestra que el niño jugó cooperativamente en un total de 67% de los intervalos de período de actividad de 60 minutos. (8/12 x 100 = 66.66%) (Cooper, Heron, and Heward, 1987).

Ejemplos de registro de muestreo de tiempo momentáneo usando intervalos de 5 minutos dentro de un período de 60 minutos (adaptado de Cooper, Heron, and Heward, 1987):

Hoja de datos del registro por intervalos

Nombre del sujeto:_____ Observador_____
Conducta_____
Fecha_____
Tiempo de inicio_____ Tiempo de término:_____
Condición:_____ Número de sesión:_____

Intervalos de 5 minutos
Presentación vertical

1	/	7	/
2	/	8	/
3	/	9	X
4	/	10	/
5	X	11	X
6	X	12	/

Intervalos de 5 minutos
Presentación horizontal

1	2	3	4	5	6
/	/	/	/	X	X
7	8	9	10	11	12
/	/	X	/	X	/

/ = Juego cooperativo
X = no juego cooperativo

Usando el registro por intervalos y el muestreo de tiempo momentáneo con varios individuos o varias conductas: frecuentemente, el analista conductual requiere registrar la misma conducta en varios individuos, en un grupo, o de varias conductas del mismo individuo (Cooper, Heron, and Heward, 1987).

Ejemplo de la hoja de intervalos de varias conductas (adaptado de Cooper, Heron, and Heward, 1987).

Sujetos:_____ Observador:_____
Conductas:_____
Fecha:_____
Tiempo de inicio:_____ Tiempo de término:_____
Condición:_____ Número de sesión:_____

Número de Intervalos	1	2	3	4	5	6	7	8	9	10	11	12
Tamaño del Intervalo	6"	6"	6"	6"	6"	6"	6"	6"	6"	6"	6"	6"
En la tarea	/			/		/	/	/				
Fuera de la tarea verbal									/			
Fuera de la tarea motora												
Fuera de la tarea pasiva		/	/							/	/	

El límite respecto al número de conductas discretas o de individuos que pueden ser observado y registrados simultáneamente se determina por grado de concordancia lazado en los datos. Esta concordancia puede decrementar, si el número de individuos de conducta(s) es excesivamente grande o las conductas muy complejas. El observador podrá registrar simultáneamente de tres a cuatro categorías. Cuando este número se excede, los acuerdos podrán incrementar cuando el observador recibe más entrenamiento y desarrolla más experiencia en el registro.

Cuando se registra a varios individuos, el criterio que usan los observadores es registrar en cada intervalo a solo un individuo. Por ejemplo, la hoja de datos muestra que John no atendió en los primeros 10 segundos del intervalo, Laura atendió durante el segundo intervalo de 10 segundos, Alicia atendió durante el tercer intervalo de 10 segundos y así sucesivamente con los otros sujetos (Cooper, Heron, and Heward, 1987).

Ejemplo del registro por intervalos de varios individuos (adaptado de Cooper, Heron, and Heward, 1987).

Observador:_____
Conducta:_____
Fecha:_____
Tiempo de inicio:_____ Tiempo de término:_____
Condición:_____ Número de sesión:_____

Número de Intervalos	1	2	3	4	5	6
Tamaño del Intervalo	6"	6"	6"	6"	6"	6"

Número de Observación x sujeto	John (10")	Laura (10")	Alicia (10")	Daryl (10")	Mariana (10")	Hans (10")
1	X	/	/	X	X	/
2	/	/	X			
3						
4						

/ = Atendiendo X = No atendiendo

Otra táctica usada con el registro de intervalos de varias categorías de respuesta se refiere al Control observacional no continuo (Tawney & Gast, 1984, citado por Cooper, Heron, and Heward, 1987). Al usar este procedimiento, se observa durante el primer intervalo y se registra durante el segundo intervalo, observa nuevamente en el tercer intervalo, y se registra en el cuarto, y así (ver el siguiente ejemplo). En el ejemplo se muestra que los primeros 6 segundos fueron usados para la observación. El observador entonces usa el segundo intervalo para registrar aquello que se observó en el primero; el estudiante ha estado en la tarea. Este procedimiento de [observa-entonces-registra] continua a lo largo del período de observación.

Ejemplo del registro por intervalos con control observacional no continuo (adaptado de Cooper, Heron, and Heward, 1987).

Observador:_____
Conducta:_____
Fecha:_____
Tiempo de inicio:_____ Tiempo de término:_____
Condición:_____ Número de sesión:_____

Número de Intervalos	1	2	3	4	5	6
Tamaño del Intervalo	6"	6"	6"	6"	6"	6"
En la tarea		/				/
Fuera de la tarea verbal				/		
Fuera de la tarea motora						
Fuera de la tarea pasiva						

Otro tipo de registro de intervalos es el procedimiento de muestreo donde se contabilizan grupos de conductas: PLACHECK - Planear la actividad a realizar - (Doke & Risley, 1972 citado por Cooper, Heron, and Heward, 1987). Con este tipo de registro se observa a un grupo de individuos al final de un período específico de tiempo contando el número de individuos ocupados en las conductas específicas comparándolas con el número total de individuos en el grupo (Cooper, Heron, and Heward, 1987).

En lo general el registro de intervalos presenta dos ventajas: 1. el registro genera un estimado de la frecuencia y de la duración de la conducta, 2. También proporciona un estimado de la conducta a lo largo del intervalo. Este registro y sus variantes requieren que el observador exclusivamente registre el comportamiento seleccionado.

Métodos de observación y registro para secuencias conductuales

Generalmente, el analista de la conducta está interesado en los patrones conductuales típicos desplegados por el individuo como tal, y sus cambios ante cierto tipo de señales y consecuencias (es decir sus posibles determinantes). La justificación de la selección del sistema de registro que el analista de la conducta lleva a cabo estriba, precisamente, en este objeto de estudio.

Estas relaciones entre la respuesta (ejecución del organismo) y el medio ambiente (señales y consecuencias) llevadas a situaciones registradas o de laboratorio, se caracterizan por ser: 1. Fracciones relativamente simples, de carácter funcional, unidireccional, y molecular (Espinosa, 1995).

No obstante, hay relaciones más complejas que se caracterizan por ser: bidireccionales, de carácter funcional y molar. Ejemplo de ello es la interacción social entre organismos pertenecientes a una especie social (Cairns, 1979; Espinosa, 1995; Hinde, 1974; Lytton, 1980). Este tipo de relaciones requieren de una descripción del total del flujo conductual, las circunstancias donde emergen y el contexto donde se desarrollan (Santoyo, 1990) por esta razón, se elaboran sistemas de observación y registro que permitieran recabar este tipo de información.

El tipo de sistema a desarrollar implica la compaginación de varios sistemas como el registro de eventos y el registro de intervalos. Del registro de eventos se toma el inicio y fin de una conducta registrada a lo largo de intervalos de tiempo continuos. La conceptualización aquí implica registrar repetidamente el flujo conductual a lo largo de un período o sesión de observación. Un ejemplo de un sistema de observación de este tipo es el SOC-IS (Santoyo y Espinosa, 1987; Santoyo, Espinosa y Bachá, 1994). Típicamente es considerado como un registro de intervalo (Anguera, 1983) pero también se registra inicio t término de las conductas. Otra característica del sistema es que las categorías conductuales son exhaustivas y excluyentes (Bakeman y Gottman, 1989; Santoyo y Espinosa, 1987).

Ejemplo de un sistema de observación conductual (adaptación de Espinosa y Santoyo, 1994).

Observador: Ma Teresa Sujeto: Matías Fecha: 04/ene/94
Escenario: Salón Nivel: preesc. "B" Sesión: 02
Condición: Línea Base 1 Total crudo: 120
H.I. 11:25 H.T. 11:35 Acuerdos: 108
Desacuerdos: 12
Concordancia: 90%

INTERVALOS DE 5 SEGUNDOS												MINUTOS
1	2	3	4	5	6	7	8	9	10	11	12	
5"	10"	15"	20"	25"	30"	35"	40"	45"	50"	55"	60"	
ac	ac	ja	ja	jp	jp	Ac	ac	des	des	ac	ac	1
												2
												3
												4
												5
												6
												7
												8
												9
												10

Observación: _____

_____

ac = actividad académica jp = juego paralelo
ja = juego aislado des = desplazamiento

Los sistemas de observación y registro mixtos permiten colectar los datos conductuales tal como ello ocurren a lo largo de los intervalos; por lo que se puede registrar tanto la frecuencia como la duración de los mismos. Además, se identifica el flujo conductual y se podrán seccionar en antecedentes y consecuentes, permitiendo el análisis secuencial de los mismos, de esta manera se hace factible identificar "estados o patrones conductuales" (Bakeman y Gottman, 1989).
En la siguiente sección se tratará tanto la concordancia como la confiabilidad entre observadores para validar la recogida de dato observacional.

La fiabilidad del dato observacional

En la literatura se utiliza con mayor frecuencia el término de confiabilidad de los datos conductuales como la concordancia entre observadores. No obstante pueden querer decir cosas diferentes. La confiabilidad se refiere a que tanto los datos obtenidos representan a las conductas observadas, es decir qué tan validos son los datos. En tanto que, la concordancia se refiere al grado de acuerdo que hay entre dos observadores o jueces sobre las conductas registradas de manera independiente (Bakeman y Gottman, 1989).

Si bien, ambos términos se encuentran estrechamente ligados, ya que uno se deriva del otro. Entonces, el primer paso es obtener la concordancia entre observadores, la que se realiza a partir de que los datos observacionales han sido recolectados. A fin de validar la concordancia entre observadores, ya sea cuando los observadores trabajan en parejas, o cuando se utilizan jueces para cotejar las conductas en video-cintas o audio-cassettes, los datos de cada observador son comparados entre si a fin de obtener el grado de acuerdo en los puntajes o datos anotados en la hoja de registro (Bakeman y Gottman, 1989).

Comúnmente, esta contrastación se hace a través de computar el número de acuerdos (Na), el número de desacuerdos (Nd), a fin de procesarse el porcentaje de acuerdos (Pa), según como queda representado en la siguiente ecuación:

$$Pa = \frac{Na}{Na + Nd} \times 100$$

Según el sistema de registro se podrán diseñar protocolos estándar para cotejar estos acuerdos y desacuerdos. En el ejemplo de registro por intervalos de secuencias conductuales (página 30) hay renglones de menor dimensión ubicados inmediatamente después de cada renglón de mayor dimensión, en este tipo de registro los desacuerdos se asientan en el renglón menor. Por lo que al sobreponer las hojas de registro de ambos observadores es posible detectar el número de desacuerdos. Esta hoja está diseñada no sólo para obtener los datos observacionales sino también como un protocolo para obtener la concordancia entre observadores.

De los puntajes anotados en esta hoja de registro y empleando la ecuación anterior tendríamos un total de 120 ocurrencias de diversas conductas, 108 acuerdos, 12 desacuerdos y una concordancia de 90%.

$$Pa = \frac{108}{108 + 12} = \frac{108}{120} \times 100 = 90\%$$

Como aclaración, si el sistema de registro del dato observacional no contiene un protocolo para obtener la confiabilidad se recomienda generar alguno. Para ello, el investigador habrá de tomar en cuenta el sistema de registro usado, las unidades conductuales, y aspectos de sincronía en el inicio de la toma de los datos observacionales para que, en efecto, se contrasten pares de registros de los datos conductuales (Espinosa y Santoyo, 1994).

Si bien, el método para obtener la concordancia del porcentaje de acuerdos de los datos observacionales es el más usado, este método presenta varias restricciones. Uno de los problemas más comunes para no tener acuerdo entre observadores es el uso de catálogos que contengan un amplio número de categorías conductuales. Un segundo problema se refiere a la complejidad de las categorías conductuales. El tercer obstáculo es que los observadores obtengan acuerdos debido al azar. En el primero y segundo problema, se minimizan cuando se usa un número reducido de categorías y códigos; además, estas categorías serán excluyentes y exhaustivas (Bakeman y Gottman, 1989).

En el tercer problema, se recomienda el uso de otro estimador de la concordancia entre observadores. Este estadístico se deriva de una matriz de confusión que es útil para controlar acuerdos debido al azar. En la siguiente matriz de confusión (Bakeman y Gottman, 1989) se coloca un ejemplo de los datos conductuales obtenidos por dos observadores.

Ejemplo de una Matriz de Confusión (adaptada de Bakeman y Gottman, 1989).

OBSERVADOR."B"

OBSERVADOR."A"	ac	Ja	jp	jg	
ac	7	0	1	0	8
ja	2	25	0	0	27
jp	0	1	25	2	28
jg	0	0	1	36	37
	9	26	27	38	100

ac = actividad académica

ja = juego aislado

jp = juego paralelo

jg = juego grupal

Bakeman y Gottman (1989) describen como se aplica el estadístico de concordancia entre observadores que corrige el azar, al que se le denomina como el coeficiente Kappa de Cohen que se deriva a través de la siguiente ecuación:

$$K = \frac{Po - Pc}{l - Pc}$$

Po es la concordancia por porcentaje obtenida de los datos conductuales reales, Pc es la proporción esperada de azar.

Po se calcula sumando los puntajes distribuidos a lo largo de la diagonal, que representa los acuerdos entre observadores (porcentaje de acuerdos).

$$Po = \frac{7 + 25 + 25 + 36}{100} = \frac{93}{100} = 0.93$$

Pc se calcula a través de multiplicar los puntajes marginales de la columna correspondiente al observador "A" contra los puntajes marginales del renglón del observador "B". En la matriz, el observador "A" registró un total de 8 intervalos donde ocurrió la actividad académica, mientras que el observador "B" registró 7 y una ocurrencia de juego paralelo. Así se continua con cada puntaje marginal de cada categoría cada par de puntajes marginales se suma con los siguientes ares. En nuestro ejemplo tendríamos:

$$Pc = \frac{9 \times 8 + 26 \times 27 + 27 \times 28 + 38 \times 37}{100 \times 100} =$$

$$Pc = \frac{72 \times 702 \times 756 \times 1406}{100 \times 100} = \frac{7406}{10000} = .7406$$

Sustituyendo ambos puntajes en el coeficiente kappa se tendría lo siguiente:

$$K = \frac{.93 - .7406}{1 - .7406} = \frac{0.1894}{0.2594} = 0.7301$$

Para interpretar los índices de concordancia entre observadores obtenida por el porcentaje de acuerdos como por el coeficiente kappa se toman los siguientes niveles. En el caso de la concordancia de acuerdos entre observadores, para el analista conductual el nivel mínimo requerido es del 80%, lo que significa que se permite no concordancia solo en un 20%, en tanto que en el coeficiente kappa se tiene tres niveles. Fleiss, (1981, citado por Bakeman y Gottman, 1989) indica que kappas regulares oscilan entre 0.40 a 0.60; buenos entre 0.60 a 0.75 y excelentes por encima de 0.75.

Compaginación de los métodos según el objeto de estudio

Generalmente, el uso de los sistemas de observación y registro de conductas se diseñaron como una herramienta metodológica que permitió obtener información sobre tipos de conductas donde no era posible utilizar sistemas más precisos de colección y medición de las mismas.

Además, la metodología observacional proporciona una excelente vía de acceso a escenarios no restringidos tal como el hábitat natural del organismo, obteniendo información confiable y valida sobre esos aspectos en que el analista conductual se encuentra interesado en estudiar.

Si el interés es el estudio de todo el flujo conductual y sus secuencias, el sistema de observación conductual que emplea el registro por intervalos resulta idóneo. Si por otra parte, el interés es estudiar que tan rápido responde un deportista al disparo en una competencia deportiva, el registro a utilizar es el de latencias. Si un investigador se encuentra interesado en el tiempo ocupado por un episodio agresivo y las veces que este comportamiento se presenta a lo largo de un período predeterminado, tendrá que elaborar un sistema de observación mixto que le den oportunidad de recolectar este tipo de información.

Por otra parte, la metodología observacional es una excelente herramienta que puede sustentar a la metodología experimental cuando se pretende que la investigación básica y no sólo la investigación aplicada se lleve a cabo in situ.

Referencias

Altmann, J. (1974) Observational study of behavior: Sampling methods. Behaviour, 49, 227-265.

Anguera, M. T. (1983) Manual de prácticas de observación, México: Trillas.

Anguera, M. T. (1991) Metodología Observacional en la Investigación Psicológica: Vol. Fundamentación (1) Barcelona: PPU.

Bakeman, R. y Gottman, J. M. (1989) Observación de la interacción: introducción al análisis secuencial. Madrid: Ediciones Morata, S. A.

Blurton Jones, N. (1980) Ethological Studies ok Child Behavior. Cambridge: Cambridge University Press.

Bunge, M. (1975) La investigación científica: su estrategia y su filosofía. Barcelona: Editorial Ariel.

Cairns, R. B. (1979) Social Interaccional Methods: An Introduction. En R. B. Cairns (Ed) The analysis of social interactions methods, issues and illustrations. New Jersey: Lawrence Erlabum Associates Publishers. 1-11.

Catania, A. C. (1974) Investigación Contemporánea en Conducta Operante. México: Trillas.

Castro, L. (1977) Diseño Experimental sin estadística: Usos y restricciones en su aplicación a las ciencias de la conducta. México: Trillas.

Cooper, J. O., Heron, J. E. y Heward, W. L. (1987) Applied Behavior Analysis. Columbus Ohio: Merrill Publishing Co.

Espinosa, M. C. y Santoyo, C. (1994) Sistema de Observación Conductual de las Interacciones Sociales: Instructivo para Observadores y Catálogo de Ejemplos. Publicación Interna del Departamento de Análisis Experimental de la Conducta. Facultad de Psicología: UNAM.

Espinosa, M. C. (1995) El estudio de la organización de las preferencias sociales en niños preescolares. Tesis de maestría inédita. México: Facultad de Psicología. UNAM.

Espinosa, M. C. (en proceso) Sistema de Observación Conductual para escenarios escolares.

Kratochwill, T. R. (1978) Singles sejetts research: strategies for evaluating change. New York: Academic Press, pp 31-68.

Hinde, R. A. (1979) Bases Biológicas de la conducta social humana. México: Siglo Veintiuno Editores.

Lytton, H. (1980) Parent-Child Interaction: The Socialization Process Observed in Twin and Singlenton Families. New York: Plenum Press.

Ribes, E. (1972) Técnicas de modificación de conducta: su aplicación al retardo en el desarrollo. México: Trillas.

Rosembluth, A. (1971) El método científico. México: Centro de investigación y de estudios avanzados del Instituto Politécnico Nacional.

Santoyo, C. y Espinosa, M. C. (1987) Un sistema de observación conductual de interacciones sociales. Revista Mexicana de Análisis de la Conducta. Vol. 13, Núms 1 y 2, 235-253.

Santoyo, C. y López, F. (1990) Análisis Experimental del Intercambio Social. México: Trillas.

Santoyo, C. (1994) Contexto e Interacción Social: Bases conceptuales y metodológicas. Barcelona: PPU.

Santoyo, C., Espinosa, M. C. y Bachá, G. (1994) Extensión del sistema de observación conductual de las interacciones sociales: calidad, dirección, contenido, contexto y resolución. Revista Mexicana de Psicología, Vol. 11, Número 1, 55-68.

Seitz, V. (1983) Methodology.

Sidman, M. (1960) Tactics of scientific research: Evaluating Experimental data in Psychology. New York: Basic Books, Inc., Publishers.

Sidman, M. (1974) algunas propiedades del estímulo de aviso en la conducta de evitación. En A. C. Catania (ed.) Técnicas de modificación de conducta: su aplicación al retardo en el desarrollo. México: Trillas. 273-280.

Skinner, B. F. (1970) Ciencia y conducta humana (una Psicología Científica). Barcelona: Editorial Fontanella. Primera edición en castellano.

Skinner, B. F. (1972) Cumulative Record: A seleccion of papers. Third edition. New York: Appleton-Century-Crofts.

Skinner, B. F. (1979) La conducta de los organismos. Barcelona: Editorial Fontanella. Segunda edición en castellano.

Smith, P. K. & Connolly, K. J. (1980) The ecology of preschool behavior. Cambridge. Cambridge University Press.

Wartofsky, M. W. (1978) Introducción a la filosofía de la ciencia. Madrid: Alianza Universidad.

El material didáctico *Metodología Observacional* fue editado originalmente por la Facultad de Psicología de la UNAM y se terminó de imprimir en febrero de 1997 en Ruesga Impresores, S. A. de C. V. 4ta. Cda. de Hidalgo No. 5, Constitución de 1917, México, D. F. Su composición se hizo en tipo Courier 12 pts. N y B. La edición consto de 600 ejemplares.

Albert Gómez, M. J. (2007). *La Investigacion educativa:claves teóricas.* España: McGraw-Hill. Pp. 99-135

CAPÍTULO 4
INSTRUMENTOS Y RECOGIDA DE DATOS DESDE EL ENFOQUE CUANTITATIVO

1. Medir desde el enfoque cuantitativo

2. Requisitos de un instrumento de medición

3. Escalas para medir aptitudes.
   - 3.1. Definición, componentes y propiedades
   - 3.2. Tipos de escalas
   - 3.3. Escala tipo Likert
   - 3.4. Escala tipo Thurstone
   - 3.5. Escala de Guttman
   - 3.6. El diferencial semántico de Osgood

4. El cuestionario
   - 4.1. Definición
   - 4.2. Preguntas
   - 4.3. Elaboración del cuestionario
   - 4.4. Aplicación del cuestionario

5. La entrevista
   - 5.1. Definición, ventajas y desventajas
   - 5.2. Tipos de entrevistas
   - 5.3. Reactivos o preguntas
   - 5.4. Fases de la entrevista

6. Pruebas e inventarios estandarizados

1. MEDIR DESDE EL ENFOQUE CUANTITATIVO

Recolectar datos implica tres actividades estrechamente vinculadas entre sí:

- Seleccionar un instrumento o método de recolección de datos entre los disponibles en el área de estudio en el cual se enmarque nuestra investigación. Este instrumento debe ser válido y fiable, de lo contrario no podremos basarnos en sus resultados.
- Aplicar ese instrumento o método de recolección de datos.
- Preparar las observaciones, registros y mediciones obtenidas para ser analizadas.

Para recolectar o recoger datos, el investigador dispone de una gran variedad de técnicas o instrumentos. En el enfoque cuantitativo, se suele utilizar un instrumento que mida las variables de interés, y medir bajo esta perspectiva consiste en asignar números a objetos o eventos de acuerdo con reglas. En las ciencias sociales, medir es el proceso de vincular conceptos abstractos con indicadores empíricos, el cual se realiza mediante un plan explícito y organizado para clasificar los datos disponibles en términos del concepto que el investigador tiene en la mente.

En el enfoque cuantitativo, aplicamos un instrumento para medir las variables contenidas en las hipótesis. Esta medición es efectiva cuando el instrumento de recolección de datos representa a las variables que tenemos en mente. Si no es así, nuestra medición es deficiente; por tanto, la investigación no es digna de tenerse en cuenta. Hay variables que son difíciles de medir, como, por ejemplo, la motivación, la inteligencia emocional; pero aun así el instrumento de medida debe acercarse lo más posible a estas variables y ser el adecuado. «Un instrumento de medición es adecuado cuando registra datos observables que representan verdaderamente los conceptos o las variables que el investigador tiene en mente. En términos cuantitativos: capturo verdaderamente la realidad que deseo capturar» (Hernández Sampieri y otros, 2003:345).

2. REQUISITOS DE UN INSTRUMENTO DE MEDICIÓN

Tanto desde un enfoque cualitativo como cuantitativo, la medición en general requiere que se contemple una serie de requisitos para que tal evaluación posea índices suficientes de credibilidad y operatividad, entre ellos los más significativos son la confiabilidad-fiabilidad y la validez.

Confiabilidad. Son sinónimos de este término estabilidad, fiabilidad, consistencia, reproductividad, predictibilidad y falta de distorsión. Por ejemplo, las personas confiables son aquellas cuyo comportamiento es consistente, predecible o fiable, lo que hacen mañana o 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; en unas ocasiones hacen algo y en otras algo distinto se dice que son inconsistentes. Lo mismo puede suceder con las mediciones, son más o menos variables de una situación a otra. Son estables y relativamente predecibles o son inestables y relativamente impredecibles.

La definición de confiabilidad se puede enfocar de tres maneras. Un enfoque se sintetiza con la pregunta siguiente: si se mide el mismo conjunto de objetos una y otra vez con el mismo instrumento de medición, ¿se obtendrían iguales o similares resultados? Esta pregunta se refiere al concepto de confiabilidad en términos de estabilidad, fiabilidad y predictibilidad.

Un segundo enfoque podría ser el que respondería a la siguiente pregunta: ¿las medidas obtenidas con este instrumento de medición son realmente las verdaderas? Es decir, ¿aquello que medimos, sin entrar en qué, independientemente de lo que sea, se está midiendo con precisión? Ésta es una definición de la confiabilidad desde la falta de distorsión.

El tercer enfoque haría referencia al error de medición, es decir, se puede investigar qué tanto de error de medición existe en un instrumento de medición. Los errores de medición son errores aleatorios y representan la suma de diversas causas. Entre dichas causas están los elementos comunes del azar o aleatorios, la fatiga temporal o momentánea, las condiciones fortuitas que en un momento en particular afectan al objeto medido o al instrumento de medición y otros factores que son temporales y cambiantes. Dependiendo del grado en que los errores de medición estén presentes en un instrumento de medición, éste será más o menos fiable. La confiabilidad desde este enfoque se define como la ausencia relativa de errores de medición en un instrumento de medición (Kerlinger, 1986:582).

Al hablar de mediciones, nos referimos igualmente a los resultados de un test como a la información recogida por medio de un cuestionario o una entrevista o mediante una observación. Desde el punto de vista práctico, puede considerarse la fiabilidad como la consistencia entre las puntuaciones que otorgan a una misma variable o evento diferentes evaluadores o la persistencia de las puntuaciones cuando se aplica el mismo instrumento de evaluación en diferentes momentos. Por tanto, la fiabilidad puede entenderse como la exactitud de los datos en el sentido de su estabilidad, repetición o precisión. Por ejemplo, si tenemos que medir la inteligencia de un grupo de alumnos y obtenemos unos datos y un mes después pasamos en distintas ocasiones el mismo instrumento a los mismos alumnos y los resultados son distintos, ese instrumento no sería fiable; si, por el contrario, obtenemos los mismos datos en una segunda fase pasándolo en distintos momentos, ese instrumento sí sería fiable.

Hay distintas formas de medir la fiabilidad, todas ellas utilizan fórmulas cuyos resultados pueden oscilar ente 0 y 1, donde un 0 significa nula fiabilidad y 1 máxima fiabilidad. Los procedimientos más utilizados son:

Test-retest Este método consiste en aplicar un mismo instrumento de medida a un mismo grupo de personas dos o más veces después de cierto período. Tal y como hemos dicho, si la correlación se acerca a 1, la fiabilidad es alta y al instrumento se le considera fiable; si, por el contrario, la correlación se aproxima a 0, la correlación es baja y al instrumento no se le considera fiable. Hay que tener en cuenta el período de tiempo desde que se pasa el instrumento la primera vez y la segunda, ya que si el período es largo, la variable puede ser susceptible de cambios, y si es corto, las personas pueden recordar cómo respondieron la primera vez. Es importante elegir ese período de tiempo teniendo en cuenta estos factores. 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 que se denomina test-retest y sirve para calcular la fiabilidad a través del tiempo.

Método de las dos mitades. Consiste en dividir el conjunto de ítem en dos partes y se comparan las puntuaciones de una parte y de la otra. Por ejemplo, se divide en total de ítem en pares e impares y se calcula la correlación entre ambas partes. Este método requiere sólo una aplicación del instrumento. Si el instrumento es fiable, las puntuaciones de ambas mitades suelen estar muy correlacionadas. Un individuo con una puntuación baja en una mitad debe tener una puntuación también baja en la otra mitad.

Método de las formas paralelas. En términos de prueba, esto equivaldría a crear dos formas de la prueba que serían equivalentes, pero no idénticas. No se trataría del mismo instrumento de medición, sino dos o más versiones equivalentes de éste. Las versiones son equivalentes en contenido, instrucciones, duración y otras características. Las versiones se administran a un mismo grupo de personas en un período de tiempo corto. Cada persona estaría sujeta a mediciones por medio de los dos instrumentos. Como resultado, cada persona tendría entonces dos puntuaciones y estos pares de puntuaciones serían utilizados en una fórmula para calcular la correlación. El instrumento es válido si la correlación entre las dos mitades es positiva (Kerlinger, 1986:592).

Coeficiente alfa de Cronbach. Este coeficiente requiere una sola administración del instrumento de medición y produce valores que oscilan entre 0 y 1. Su ventaja reside en que no es necesario dividir en dos mitades a los ítem del instrumento, simplemente se aplica la medición y se calcula el coeficiente.

Coeficiente KR-20 de Kurder y Richardson. Estos autores desarrollaron un coeficiente para estimar la confiabilidad de una medición cuya interpretación es la misma que la del coeficiente alfa (Hernández Sampieri y otros, 2003:354).

Otro requisito de un instrumento de medida es la validez. La validez, en términos generales, se refiere al grado en que un instrumento mide la variable que pretende me-

dir. Se puede sintetizar en la pregunta ¿estamos midiendo lo que creemos que estamos midiendo? Por ejemplo, un profesor quiere medir la comprensión lectora de sus alumnos y realiza una prueba en la que incluye sólo elementos de velocidad lectora. Esta prueba no es válida, ya que aunque quizá mida muy bien la velocidad lectora no está midiendo la comprensión, que es lo que él quiere medir. Así, una prueba de inteligencia debe medir la inteligencia y no la memoria, etc. Esto que en apariencia es fácil no lo es tanto cuando las variables a medir se complican, como puede ser la motivación, los sentimientos, las emociones. Una prueba es válida de acuerdo con el propósito científico o práctico de quien la utiliza.

La validez es un concepto del que pueden tenerse diferentes tipos de evidencia, o lo que es lo mismo, podemos señalar distintos tipos de validez: la relacionada con el contenido, con el criterio y con el constructo.

Validez de contenido. Este tipo de validez se refiere al grado en que un instrumento refleja un dominio específico de contenido de lo que se mide. Es el grado en que la medición representa el concepto medido. Se puede definir también como la representatividad o la adecuación de muestreo del contenido, sustancia, materia o tema de un instrumento de medida.

Cualquier propiedad educativa posee un universo teórico de contenido de la propiedad que mide, que consiste en todas las posibles cosas que se dicen u observan acerca de la propiedad. A cada cosa que se dice u observa de esa propiedad se le denomina reactivo. Para que un instrumento de medida posea una alta validez de contenido debe tratar todos los reactivos de esa propiedad. Por ejemplo, si el universo de una propiedad está formado por los reactivos 1; 2, 3, el instrumento debe tratarlos todos. Así, por ejemplo, en un test sobre personalidad, la prueba debe tratar todos los factores que desde un punto de vista teórico configuran la personalidad: un test de operaciones matemáticas para que tengan alta validez de contenido debe tratar todas ellas: sumas, restas, multiplicaciones, etc. Cuantos más reactivos de la propiedad a estudiar trate, más alta será la validez de contenido.

El cálculo de la validez de contenido radica principalmente en el juicio. Solo o con otros, el investigador juzga la representatividad de los reactivos; para ello, lo primero es revisar cómo ha sido medida la variable por otros investigadores y en base a esa revisión elaborar un universo de ítem posibles, consultando con investigadores familiarizados con la variable (jueces) para ver si el universo ha sido exhaustivo. La validez de contenido es cuantificable a través del empleo de índices de concordancia de las evaluaciones de los jueces; uno de esos índices puede ser el índice *Kappa* de Cohen (Cohen, 1960).

Validez de criterio. Este tipo de validez se estudia al comparar las puntuaciones de una prueba o escala con algún criterio externo. Cuanto más se relacionen los resultados del instrumento de medición con el criterio, la validez de criterio será mayor.

Podemos señalar dos tipos de validez de criterio: la predictiva y la concurrente. La característica que diferencia una de otra es la dimensión del tiempo. La *validez pre-*

dictiva fija el criterio en el futuro; por ejemplo, una prueba para determinar la capacidad de los operarios de una cadena de producción se validará comparando los resultados con el desempeño de su trabajo. Por su parte, la *validez concurrente* fija el criterio en el presente y los resultados del instrumento se correlacionan con el criterio en el mismo momento o punto de tiempo. Por ejemplo, un cuestionario para detectar las tendencias de voto sobre un tema concreto que posteriormente se podrá comparar con los resultados de la elección (Bohrnstedt, 1976). Para estimar la validez de criterio, el investigador correlaciona su medición con el criterio externo y este coeficiente se toma como coeficiente de validez.

Validez de constructo. Este tipo de validez se refiere al grado en el que una medición se relaciona de manera consistente con otras mediciones de acuerdo con hipótesis derivadas teóricamente y que conciernen a los conceptos que se están midiendo (Hernández Sampieri y otros, 2003:349). Cuando los expertos en medición investigan la validez de constructo de una prueba, 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. Su interés por lo general está centrado en las propiedades que se miden más que en las pruebas utilizadas para lograr la medición. Por ejemplo, supongamos que un investigador desea evaluar la validez de constructo de una prueba para medir el rendimiento del alumno y que el investigador considera que aspectos como la tendencia voluntarista del alumno, llamémoslo 1; la capacidad del alumno, llamémoslo 2; con la motivación intrínseca, 3, y con el estatus social, 4, son factores importantes del rendimiento y que la postura teórica, producto de otras investigaciones, fuese que ese rendimiento se correlacionara positivamente con 1, 2 y 3 y negativamente con 4, entonces ese instrumento mediría en realidad el rendimiento. Si esto fuese así, podríamos decir que ese instrumento tiene una buena validez de constructo.

El aspecto más importante de la validez de constructo, y que además la separa 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 validez de constructo incluye tres etapas

1. Se establece y especifica la relación teórica entre los conceptos.
2. Se correlacionan ambos conceptos y se analiza cuidadosamente la correlación.
3. Se interpreta la evidencia empírica de acuerdo con el nivel en el que se clarifica la validez de constructo de una medición en particular.

En la validez de constructo es importante que haya un soporte teórico, es decir, que esté fundamentado desde otras investigaciones que los conceptos están relacionados. Cuanto más elaborado y comprobado se encuentre el marco teórico que apoya la hipótesis, la validación de constructo arrojará mayor luz sobre la validez de constructo de un instrumento de medición.

Para determinar este tipo de validez suele utilizarse un procedimiento denominado «análisis de factores»; para este análisis se precisan conocimientos estadísticos y un programa como el SPSS.

Como conclusión, podemos decir que la validez es un requisito imprescindible para un instrumento de medición y que, además de ser confiable, ha de ser válido; para ello, debe tener altos coeficientes en la validez de contenido, en la validez de criterio y en la validez de constructo. Así, la validez total de instrumento ha de ser la suma de los tres tipos de validez mencionados: cuanto más altos sean cada uno de ellos, más alta será la validez total del instrumento de medida.

3. ESCALAS PARA MEDIR ACTITUDES

Existen tres tipos principales de escalas de actitud: escalas de puntuación sumada (un tipo de las cuales es la llamada tipo Likert), escala de intervalos aparentemente iguales (llamadas escalas de Thurstone) y escalas acumulativas o de Guttman.

Antes de estudiar distintas escalas para medir actitudes es importante que sepamos que es una actitud y cuáles son los elementos que la forman. Veamos cada una de estas partes

3.1. Definición, componentes y propiedades

Actitud es un estado de disposición psicológica adquirida y organizada a través de la propia experiencia que incita al individuo a reaccionar de una manera característica frente a determinadas personas, objetos o situaciones (Fernández de Pinedo, *http://.mtas.es/insht/ntp 015.htm*); es decir, si la persona hace una evaluación positiva hacia un determinado objeto, entonces su actitud hacia ese objeto es positiva o favorable, esperándose también que sus manifestaciones de conducta (respuestas) hacia dicho objeto sean en general favorables o positivas, mientras que si la evaluación es negativa o en contra del objeto, las actitudes serán negativas o desfavorables. Teóricamente, se asume que una actitud no tiene sólo una dirección, es decir, es favorable o desfavorable, sino que existen grados ubicados entre estos dos polos formando un continuo actitudinal.

Las actitudes no son innatas, sino que se forman a lo largo de la vida; no son directamente observables, sino que han de ser inferidas a partir de ciertas respuestas verbales o no verbales del sujeto. Las respuestas mensurables de la actitud se llaman componentes y son tres: *componente afectivo, componente cognoscitivo* y *componente conductual.*

Componente afectivo. Está definido por los sentimientos que el individuo tiene hacia el objeto de la actitud y la intensidad de los mismos son las sensaciones y sentimientos que dicho objeto produce en el sujeto.

Componente cognoscitivo. Viene definido por el conjunto de datos e información que el sujeto sabe acerca del objeto del cual toma su actitud. Un conocimiento detallado del objeto favorece la asociación al objeto.

Componente conductual. Son las interacciones, disposiciones o tendencias hacia un objeto. Es cuando surge una verdadera asociación entre objeto y sujeto.

Por su parte, las actitudes tienen diversas propiedades entre las que destacan la dirección (positiva o negativa) e intensidad (alta o baja), factores muy importantes a tener en cuenta a la hora de codificar las alternativas de respuesta.

Tanto el concepto de actitud como los componentes que la forman son aspectos importantes a tener en cuenta a la hora de elaborar escalas que midan las actitudes. Entre esas escalas, tenemos las escalas tipo Likert, Thurstone, Guttman y diferencial semántico.

3.2. Tipos de escalas

Existen múltiples métodos para el análisis de las actitudes al igual que existen diversas formas de concebirlas. Antes de exponer las escalas anteriormente citadas, conviene recordar los tipos de escala que miden el componente afectivo de la actitud. Según la tipología de Stevens, distinguimos cuatro tipos diferentes de escala.

Nominales. Consisten en la clasificación de algún objeto en dos o más categorías (por ejemplo, SÍ/NO). En este tipo de escala, el orden de las categorías carece de importancia, pues lo único que nos proporciona es la equivalencia de los individuos en la relación de los objetos. De este modo, no podremos diferenciar a los individuos en base al grado en que poseen un atributo, sólo sabremos si lo poseen o no. Por ejemplo, en una escala para medir la motivación (suponiendo hipotéticamente que la motivación sea escalonable nominalmente) nos diría si los individuos poseen el atributo motivación o no lo poseen, pero no en el grado en que lo poseen.

Ordinales. Recordemos que esta escala se basa en el orden de los objetos. Aunque no nos aporta ninguna idea sobre la distancia que existe entre ellos, nos permite clasificar a los individuos en función del grado en que poseen cierto atributo. Por ejemplo, con respecto a la temperatura 40 °C no el doble de 20 °C, pero sí nos indica que es una temperatura más alta. Esta escala nos permite ordenar aunque no dispongamos de una unidad de medida para saber las distancias que separan a los individuos.

De intervalo. Con esta escala sabemos las distancias, pero no el principio métrico sobre el que se han construido los intervalos, es decir, podríamos suponer que los intervalos son iguales. Por ejemplo, las distancias de un metro son centímetros iguales unos a otros, pero lo que no lograríamos sería fijar un punto cero y estar seguros de que una puntuación 2 es dos veces una puntuación 1.

De proporción. Con estas escalas logramos construir intervalos iguales y además situar un punto cero en la escala.

3.3. Escala tipo Likert

Para medir un objeto se requiere una escala de medida. Definimos una escala como una serie de ítem, entendiendo por ítem una frase o proposición que expresa una idea po-

sitiva o negativa respecto a un fenómeno que nos interesa conocer, que han sido cuidadosamente seleccionados de forma que construyan un criterio válido, fiable y preciso para medir de alguna forma los fenómenos sociales. En nuestro caso, este fenómeno será una actitud cuya intensidad queremos medir.

La escala tipo Likert fue desarrollada por Rensis Likert al principio de los años treinta, siendo una de las más utilizadas en la medición de las actitudes. Se trata de un conjunto de reactivos de actitud donde todos los reactivos son considerados con un valor de actitud aproximadamente igual y donde cada uno de los participantes señala con grados de acuerdo o desacuerdo. Está formada por un conjunto de ítem presentados en forma de afirmaciones o juicios ante los cuales los sujetos tienen que manifestarse. Se trata de una escala aditiva, lo que significa que las puntuaciones de los reactivos de dicha escala se suman para producir una puntuación de actitud del individuo. El propósito de la escala de puntuación sumada es ubicar a un individuo en algún punto del continuo del nivel de acuerdo de la actitud (Kerlinger, 2002:645).

Al interrogado se le presenta una afirmación y se le pide que extreme su reacción eligiendo uno de los cinco puntos de la escala (muy de acuerdo, de acuerdo, indeciso, en desacuerdo, muy en desacuerdo). A cada punto se le asigna un valor numérico. La suma algebraica de las puntuaciones de las respuestas del individuo a todos los ítem es su puntuación total, que se entiende como representativa de su posición favorable-desfavorable con respecto al fenómeno que se mide.

Esta escala asume un nivel de medida ordinal, de tal forma que los sujetos son ordenados en la escala en función de su posición agradable o desagradable sin que aporte ninguna idea sobre la distancia que existe entre ellos, es decir, si una persona obtiene una puntuación de 60 puntos en una escala no significa que su actitud ante el fenómeno medido sea el doble que la de otro individuo que haya obtenido una puntuación de 30, pero sí nos informa que el de 60 puntos tiene un actitud más favorable que el de 30.

La construcción de la escala comporta los siguientes pasos:

1. Definición del objeto actitudinal.
2. Determinar la categoría de los ítem.
3. Administración de la escala a una muestra representativa.
4. Análisis de los ítem.

Definición del objeto actitudinal. En este primer paso hemos de especificar muy claramente el objeto sobre el cual vamos a intentar la medida de la actitud. Tal objeto está relacionado, evidentemente, con nuestros objetivos de investigación. A través de una revisión bibliográfica, así como de la consulta de otros instrumentos, se recoge una serie de ítem relacionada con la actitud que queremos medir y se recogen aquellos que expresan una actitud claramente agradable o desagradable. No es necesario que todos los ítem estén formulados de forma positiva, sino que podrán combinarse entre positi-

Ejemplo:

Los contenidos que estudiamos en el colegio son prácticos para la vida profesional.

Los profesores no dedican tiempo a hablar con los alumnos.

Aunque no es estrictamente necesario, conviene que la mitad de los ítem sea favorable y la otra mitad desfavorable. Se aconseja presentar de forma aleatoria los ítem favorables y los desfavorables para que el efecto cercanía no influya en la elección.

El siguiente paso será determinar la categoría de los ítem.

Tal y como hemos dicho, las afirmaciones pueden tener una dirección favorable o positiva y desfavorable o negativa. Si la afirmación es positiva, significa que califica favorablemente al objeto de actitud, y cuanto más de acuerdo con la afirmación estén los sujetos, su actitud será más favorable. En este caso, la codificación sería:

(5) Muy de acuerdo.

(4) De acuerdo.

(3) Indeciso.

(2) En desacuerdo.

(1) Muy en desacuerdo.

Es decir, estar más de acuerdo implica una puntuación mayor.

Si, por el contrario, la afirmación es negativa, significa que califica desfavorablemente al objeto de actitud, y cuanto más de acuerdo estén lo sujetos con la afirmación, su actitud es menos favorable, es decir, más desfavorable. En este caso, la codificación sería:

(1) Muy de acuerdo.

(2) De acuerdo.

(3) Indeciso.

(4) En desacuerdo.

(5) Muy en desacuerdo.

El siguiente paso será la administración de la escala a una muestra representativa. Es decir, se selecciona un grupo de sujetos similar a aquel al que se piensa aplicar la escala. Éstos responden eligiendo en cada ítem la alternativa que mejor describa su posición personal. Obtendremos para cada sujeto una puntuación global que nos permita estimar la posición del sujeto en un continuo hipotético.

Análisis de los ítem. Existen dos procedimientos aplicables en esta fase:

a) Uno sería el estudio de las correlaciones entre cada puntuación total. Las correlaciones nulas o bajas nos harán prescindir del ítem, mientras que las elevadas señalan que el ítem proporciona información relevante para la conducta estudiada.

b) Otra forma sería seleccionar el 25 por 100 de los sujetos con puntuación más alta y el 25 por 100 con puntuación más baja. Se analizan si existen diferencias estadísticamente significativas entre ambos grupos para cada uno de los elementos de la escala. Si resultan diferencias significativas, el elemento se incluye; mientras que si no se dan diferencias significativas, el ítem se elimina.

Con los criterios anteriores, se selecciona el número de ítem deseado para la escala siendo conveniente, tal y como hemos dicho, que la mitad de la escala exprese una posición favorable y la otra mitad desfavorable. El número de ítem suele oscilar entre quince y treinta.

Existen dos formas de aplicar una escala tipo Likert. La primera es de manera autodeterminada: se le entrega la escala al respondiente y éste marca respecto a cada afirmación la categoría que mejor describe su reacción o respuesta. La segunda forma es la entrevista, donde un entrevistador lee las afirmaciones y alternativas de respuesta al sujeto y anota lo que éste conteste. Cuando se aplica vía entrevista, es necesario que se le entregue al entrevistado una tarjeta donde se le muestre a éste las alternativas de respuesta o categorías (Hernández Sampieri y otros, 2003:379).

3.4. Escala tipo Thurstone

Las escalas de intervalos aparentemente iguales de Thurstone logran escalar los reactivos de la actitud. A cada registro se le asigna un valor de escala que indica la fortaleza de la actitud de una respuesta de acuerdo con el reactivo. El universo de reactivos se considera un conjunto ordenado, es decir, los reactivos difieren en su valor de escala. El propósito de esta escala es escalar los estímulos que se presentan a lo largo de un continuo desde los más favorables a los menos favorables.

Para construir una escala de actitud tipo Thurstone, señalaremos los siguientes pasos:

Especificación de la variable. Se trata de especificar lo más claramente posible cuál es la variable de actitud que queremos medir. Al igual que en la escala Likert, tal objeto ha de estar relacionado con nuestros objetivos. Conviene formularlo con claridad y sin ambigüedad para obtener una buena escala que mida lo que quiere medir.

El siguiente paso será la recolección de enunciados. Se trata de recoger información para la construcción de los ítem que van a componer la escala. La escala requiere tantos ítem como sean necesarios para cubrir toda la gama, que va desde los muy desfavorables al objeto sobre el que intentamos medir la actitud hasta los muy favorables. Dichos enunciados se pueden conseguir por diferentes métodos: a través de la literatura

sobre el tema, a través de una entrevista previa a cualquier persona que consideremos representativa y la propia intuición del investigador.

La **selección de los ítem** en una lista previa es el tercer paso en la construcción de nuestra escala. Esta selección se refiere al primer filtro que deben pasar los ítem de la escala. Dicho filtro ha de tener en cuenta, sobre todo, las características lingüísticas y gramaticales de los ítem, su estructura lógica y sus características generales.

El **recurso a los jueces** es la particularidad más importante del método de intervalos aparentemente iguales para establecer el grado de favorabilidad de cada ítem dentro de la escala y, por tanto, para abandonar aquellos ítem que nos proporcionan una información redundante. Se trata de clasificar los ítem en una escala imaginaria que representa la variable actitud en función de su grado de favorabilidad-desfavorabilidad de los mismos recurriendo a terceras personas (jueces). A éstas no se les pide que manifiesten su actitud hacia los ítem que les presentamos, sino que nos digan si cada ítem muestra una tendencia favorable o desfavorable y no la actitud de la persona juez hacia el ítem.

Cálculo del valor escalar de cada ítem. El valor escalar de un enunciado en una escala Thurstone (el lugar que ocupa en el continuo de la escala) viene dado por la mediana de las respuestas de los jueces a dicho enunciado, es decir, por la medida de tendencia central que deja la mitad de los individuos de la distribución a cada lado.

El siguiente paso será la **depuración de la escala mediante el criterio de ambigüedad.** No todos los jueces colocan el enunciado en el mismo intervalo. En la medida en que los jueces concuerden más en sus colecciones, el ítem será menos ambiguo. Esto se resuelve mediante la desviación cuartil: un ítem será menos ambiguo en tanto en cuanto su desviación cuartil sea menor.

El último paso será la **selección de enunciados uniformemente distribuidos.** Si, por ejemplo, construimos una escala a partir de once intervalos, lo ideal es una escala con veintidós ítem, es decir, dos ítem por intervalo situados preferiblemente en sus límites y en su centro.

Una vez elegidos los enunciados que formarán parte de la escala, se crea la misma y se distribuye a la muestra que hemos elegido. A los sujetos sólo se les pedirá que marquen con una señal los enunciados con los que estén de acuerdo. Como nosotros conocemos ya el valor escalar de cada uno de ellos, la medida de la actitud de cada individuo será la medida de los valores escalares de los ítem con los que está de acuerdo (Elejarrieta-Íñiguez-Tramunt, 1984).

3.5. Escala de Guttman

La escala acumulativa o de Guttman consta de un conjunto relativamente pequeño de reactivos homogéneos que son unidimensionales. Una escala unidimensional mide una variable y sólo una. La escala obtiene su nombre de la relación acumulativa entre los reac-

tivos y las puntuaciones totales de los individuos. Esta escala se basa en el principio de que algunos ítem indican en mayor medida la fuerza o intensidad de la actitud.

Hemos visto cómo la escala de Likert situaba a un individuo dentro de un punto de un continuo; la escala de Thurstone lo que sitúa en el continuo es a un reactivo; el modelo de Guttman permite escalar a los sujetos y a los estímulos o reactivos conjuntamente.

La idea en la que se basó Guttman fue en la posibilidad de ordenar los estímulos, de tal forma que si un sujeto responde correctamente a uno de ellos, lo habrá hecho también a todos los que estén situados por debajo de dicho estímulo en la escala resultante. Es decir, si en una escala acumulativa de diez ítem el encuestado marca un 4, eso debería significar su acuerdo con las cuatro primeras afirmaciones; si marca un 8, debería estar de acuerdo con las ocho primeras. A esto es a lo que se le llama *modelo ideal*. La finalidad es encontrar un grupo de ítem que corresponda con este esquema. En la práctica, rara vez encontraremos este esquema acumulativo en su forma perfecta; por tanto, usamos el análisis de escalograma para examinar cómo de estrechamente un grupo de ítem se corresponde con esta idea de acumulatividad ideal. Un ejemplo de escala acumulativa ideal sería la siguiente (+ significa favorable y – desfavorable):

	Elementos					
Sujetos	1	2	3	4	5	Puntuación
A	+	+	+	+	+	5
B	+	+	+	+	–	4
C	+	+	+	–	–	3
D	+	+	–	–	–	2
E	+	–	–	–	–	1
F	–	–	–	–	–	0
Puntuación ítem	5	4	3	2	1	

El sujeto que marca positivo el elemento 3, también lo hace en los inferiores: 2, 1; el que marca el 4, lo hará también en el 3, 2, 1; el que marca el 5, lo hará en el 4, 3, 2, 1, etc.

El problema está en determinar cuál es el margen de desviación permitida para que se puedan aceptar los datos obtenidos. Se considerará error cada una de las desviaciones del patrón de respuestas obtenidas con respecto al modelo ideal; por ello, lo que hay que determinar es si el número de errores es lo suficientemente bajo como para garantizar que la escala cumple las condiciones del modelo acumulativo ideal (De Lara Guijarro-Ballesteros Velázquez, 2001:345).

Un ejemplo de error con respecto a la escala ideal sería el siguiente:

	Elementos					
Sujetos	1	2	3	4	5	Puntuación
A	+	+	—	+	+	5
B	+	—	+	+	—	4
C	+	+	+	—	—	3
D	+	+	—	—	+	2
E	+	—	—	—	—	1
F	—	—	—	—	—	0
Puntuación ítem	5	3	2	2	2	

Según la escala ideal, si el sujeto A contesta afirmativamente al elemento 5 debería contestar también afirmativamente a los elementos 4, 3, 2, 1; en la matriz podemos comprobar que no es así, a esto es a lo que se le considera un error. Así, el sujeto A tiene un error; el B tiene dos errores; el C no tiene ningún error, por lo que se ajusta a la escala ideal; el D tiene dos errores; el E y el F tampoco tienen errores, también se ajustan a la escala ideal. El total de errores en la escala es de 5.

Uno de los procedimientos para controlar el margen de desviación permitida con respecto a la escala ideal es el coeficiente de reproductividad; dicho coeficiente establece una proporción entre el número de errores y el número total de respuestas. Si el coeficiente es igual o superior a 0,90, se considera que la escala cumple los requisitos de la escala acumulativa perfecta.

Para construir la escala como en las anteriores, comenzaremos por:

Definir el objeto, que ha de estar de acuerdo con la investigación, por ejemplo, la igualdad entre los sexos.

-**Desarrollo** de una larga lista de ítem que reflejen el concepto.

Valoración de un grupo de jueces en términos de cómo de favorables son con el concepto que hemos elegido: ellos darían «sí» si el ítem supone una actitud favorable hacia la igualdad entre sexos y un «no» si no es así. No se trata de las creencias de los jueces sobre el tema, sino que lo que deben juzgar es la propia afirmación en relación a la cuestión objeto de estudio.

Desarrollo de la escala acumulativa. Para ello, debemos construir una matriz o tabla que muestre las respuestas de todos los que han contestado a todos los ítem. Se debe ordenar la matriz de forma que los que hayan manifestado su acuerdo con un mayor nú-

mero de afirmaciones aparezcan en la parte superior y los que hayan manifestado mayor número de desacuerdos aparezcan en la parte inferior.

Administración de la escala. Una vez seleccionados los ítem, se le presenta al encuestado y se le solicita que señale aquellos con los que está de acuerdo. Cada ítem de la escala debe tener asociado el valor obtenido a partir del análisis del escalograma. Para computar la puntuación del encuestado sumamos los valores de dada uno de los ítem con los que manifestó estar de acuerdo.

3.6. El diferencial semántico de Osgood

El diferencial semántico fue desarrollado por Osgood y Tannenbau (1957) para explorar las dimensiones del significado; hoy día consiste en una serie de adjetivos extremos que califican al objeto de actitud ante los cuales se solicita la reacción del sujeto. Éste debe calificar al objeto de actitud con un conjunto de adjetivos bipolares; entre cada par de adjetivos se presentan varias opciones y el sujeto selecciona aquella que en mayor medida refleje su actitud.

Ejemplo :

Dulce _____ _____ _____ _____ _____ _____ _____ Amargo

Los adjetivos son extremos y entre ellos hay siete opciones de respuesta. Si el sujeto considera que su actitud está muy estrechamente relacionado con los extremos (*MR*), señalará con *X* la posición más cercana al extremo correspondiente; si lo considera estrechamente relacionado (*R*), señalará el tramo siguiente, dependiendo del extremo; si lo considera mediano (*M*), será el tramo siguiente, y si lo considera medio, la señal ocuparía una posición neutral (*N*), igual en ambas direcciones.

Dulce	*X*						Amargo
	MR	*R*	*M*	*N*	*M*	*R*	*MR*
	7	6	5	4	3	2	1

La cuantificación de la escala puede ser de distintas formas. Una de ellas puede ser del 1 al 7, sabiendo que el 1 corresponde a un extremo y el 7 al otro extremo, y otra de –3 a +3.

Dulce							Amargo
7	6	5	4	3	2	1	
3	2	1	0	–1	–2	–3	

La puntuación del 1 al 7 es la más utilizada, ya que evita trabajar con números negativos.

Para la construcción de la escala debemos, al igual que en las escalas anteriores, identificar el objeto de medición de la actitud siguiendo los mismos pasos que en las escalas anteriores.

El primer paso será generar una lista de adjetivos bipolares exhaustiva y aplicable al objeto de actitud a medir. No hay un límite exacto de cuantos adjetivos deberán conformar la escala. El investigador deberá decidir teniendo en cuenta la suficiencia y la representatividad del contenido.

Construir una versión preliminar de la escala y la administramos a un grupo de sujetos a manera de prueba piloto.

Correlacionar las respuestas de los sujetos para cada par de adjetivos o ítem: se ha de correlacionar un ítem con todos los demás (cada par de adjetivos contra el resto).

Calcular la confiabilidad y la validez de la escala total (todos los pares de adjetivos).

Seleccionar los ítem que presenten correlaciones significativaś con los demás ítem. Se seleccionarán aquellos en los que la correlación sea significativa.

Desarrollar la versión final de la escala. Se confecciona la escala final con aquellos que tienen una correlación significativa.

Su interpretación depende del número de ítem o pares de adjetivos. En ocasiones, se califica el promedio obtenido en la escala total:

$$\frac{\textit{Puntuación total}}{\textit{Número de ítem}}$$

4. EL CUESTIONARIO

4.1. Definición

Una de las técnicas de recogida de datos más usual es el cuestionario. Se le puede definir como una técnica estructurada que permite la recogida rápida y abundante de información mediante una serie de preguntas orales o escritas que debe responder un entrevistado con respecto a una o más variables a medir.

El cuestionario ha de cumplir la función clave de servir de nexo de unión entre los objetivos de la investigación y la realidad de la población encuestada. Por tanto, el cuestionario deberá, por una parte, traducir en sus preguntas los objetivos de la investigación, y por otra, suscitar en los encuestados respuestas sinceras y claras cuya información podrá ser clasificada y analizada posteriormente (De Lara Guijarro-Ballesteros Velázquez, 2001).

Los cuestionarios pueden ser monotemáticos o politemáticos, según intenten cercar un objeto de estudio desde una o varias problemáticas con respecto a una variable o varias variables a medir. Tiene como ventajas la rapidez, la facilidad de aplicación y la posibilidad de ser constatado por muchos sujetos. Los inconvenientes vienen dados por la falta de sinceridad, la adecuación al léxico, la superficialidad y la concordancia de las respuestas en las preguntas abiertas.

4.2. Preguntas

Cada cuestionario obedece a diferentes necesidades y problemas de investigación lo cual origina que en cada caso el tipo de pregunta sea diferente. Básicamente, se puede considerar dos tipos: abiertas y cerradas.

Las *preguntas abiertas* no delimitan de antemano las alternativas de respuesta, por lo cual el número de categorías de respuesta puede ser muy elevado, ya que el sujeto puede escribir lo que quiera. Este tipo de preguntas se utiliza principalmente cuando no se tiene información sobre las posibles respuestas de los sujetos, cuando esta información es insuficiente o cuando queremos profundizar en una opinión. Presentan algunas desventajas, como las relacionadas con la codificación, clasificación, preparación para el análisis, así como la dificultad que representan para aquellas personas que no dominan el lenguaje oral y escrito; el tiempo de respuesta es más largo, por lo que requieren un mayor esfuerzo y tiempo.

Un ejemplo de pregunta abierta sería:

*¿Cuales son los motivos que le empujan a estudiar?*_____

_____

*¿Qué opina sobre la educación a distancia?*_____

_____

Las preguntas abiertas se codifican una vez que conocemos todas las respuestas de los sujetos a los que se les pasó el cuestionario o al menos las principales tendencias de respuestas. El procedimiento consiste en encontrar y darles nombre a los patrones generales de respuesta, listar esos patrones y asignarles un valor numérico.

Las *preguntas cerradas* contienen categorías o alternativas de respuesta que han sido delimitadas y codificadas previamente. Es decir, se presentan a los sujetos posibles respuestas y quien responde debe ceñirse a ellas. Pueden ser dicotómicas o con varias alternativas.

Ejemplos:

¿Pones interés en los estudios?

(SÍ)

(NO)

¿Cuántas horas dedicas a estudiar a la semana?

1 a 2 horas

2 a 3 horas

3 a 4 horas

más

A las preguntas cerradas, al contrario de las abiertas, son fáciles de asignar valores numéricos o codificar y preparar para su análisis, requieren un menor esfuerzo y tiempo para el sujeto, ya que no tienen que escribir y verbalizar la respuesta, sino elegir una de las que se les presentan. Al igual que las respuestas abiertas, las cerradas tienen desventajas: la más importante es que limitan las respuestas de la muestra y en ocasiones ninguna de las categorías responde con exactitud a lo que las personas tienen en mente (Hernández Sampieri, 2003).

Para formular este tipo de pregunta es necesario tener en cuenta varios aspectos:

- En primer lugar, se ha de tener en mente las posibles respuestas del sujeto.
- Se ha de asegurar que el sujeto comprende las categorías de las respuestas.
- Deberán aparecer todas las opciones posibles de respuestas.
- Han de estar redactadas con claridad, de forma que no lleven a equívocos o a ambigüedades.

La elección de un tipo u otro de pregunta, tal y como hemos dicho, va a depender del grado en que se puedan anticipar las posibles respuestas, los tiempos de que se disponga para codificar y si se requiere una respuesta más precisa o profundizar en alguna cuestión.

Independientemente de que las preguntas sean abiertas o cerradas y de que estén codificadas o no, todas deben tener las siguientes características:

- Las preguntas han de ser claras y comprensibles para los respondientes, deben evitarse términos confusos o ambiguos.
- Las preguntas no deben incomodar al respondiente. Cuando se trata de preguntas personales, el sujeto puede sentirse incómodo y es mejor hacerlas de forma sutil.
- Las preguntas han de referirse preferentemente a un solo aspecto o a una relación lógica. Por ejemplo, a la pregunta de si usted va de vacaciones a la playa o a la montaña puede ser motivo de confusión: por tanto, es mejor dividir la pregunta en dos aspectos.
- Las preguntas no deben inducir a la respuesta, se han de evitar preguntas tendenciosas o que den pie a elegir un tipo de respuesta.
- Las preguntas no pueden apoyarse en instituciones, ideas respaldadas socialmente ni en evidencia comprobada
- El lenguaje utilizado en las preguntas ha de ser el apropiado a las características del respondiente.
- Las preguntas deberán ser breves en lo posible con el fin de evitar interpretaciones difíciles.
- Han de ser formuladas en forma neutral o, como mucho, en positivo, nunca en negativo.

- Emplear la forma personal y directa
- Se ha de tener en cuenta a quién va dirigida.

Otro aspecto importante en el cuestionario es la secuencialidad de las preguntas: se pueden ordenar de distinta forma y seguir diversos criterios.

En algunos casos, es conveniente empezar con preguntas neutrales o fáciles de contestar para que el sujeto vaya adaptándose al cuestionario, dejando para el final las más delicadas.

- En ocasiones, las cuestiones identificativas se incluyen al principio, pero en otras se ponen al final, sobre todo cuando los sujetos puedan sentir que se comprometen si responden al cuestionario.
- La colocación, conocida por técnica del embudo o de concentración, es otra forma de ordenar las preguntas. Esta técnica lo que hace es que presenta primeramente las preguntas más generales para ir poco a poco a las particulares o específicas, y la del embudo invertido donde las preguntas van de lo más específico a los temas más generales.
- La técnica de la dispersión es otra forma de ordenar las preguntas y consiste en espaciar aquellas preguntas afines con la intención de que unas respuestas no influyan en las otras.

El tamaño del cuestionario es otro factor importante a tener en cuenta. El cuestionario debe tener las preguntas justas y necesarias para realizar el trabajo. Existe una tendencia a formular excesivas cuestiones que posteriormente no son utilizadas. Un cuestionario excesivamente largo tiene el inconveniente de cansar a los encuestados y puede inducir a no responderle. Por otra parte un cuestionario excesivamente corto corre el riesgo de perder información.

El diseño del cuestionario es algo que no debemos olvidar, ya que la imagen del mismo será nuestra imagen y la seriedad con la que nos respondan dependerá en parte de la seriedad del cuestionario.

En primer lugar, el cuestionario debe tener una portada que recoja el título del mismo, la autoría y el propósito del mismo, así como algún otro elemento que lo caracterice: empresa, fecha, a quién va dirigido...

En segundo lugar, debe aparecer una pequeña explicación y las instrucciones para su cumplimentación. Las instrucciones son muy importantes dado que es el medio de evitar la introducción de elementos subjetivos en las respuestas. En ellas se expondrán las advertencias sobre la forma de cumplimentar el cuestionario. Es necesario que sean claras para los usuarios a quienes van dirigidas.

En ocasiones, puede ser interesante adjuntar una carta dirigida al encuestado solicitando su cooperación garantizándole la confidencialidad y dándole las gracias por su ayuda.

4.3. Elaboración del cuestionario

En este apartado seguiremos a De Lara Guijarro y Hernández Sampieri.

Las fases para la elaboración del cuestionario son:

1. Definición de los objetivos de estudio.
2. Determinación de las variables a tratar.
3. Revisión de la literatura de cuestionarios que midan las mismas variables que pretende medir la investigación.
4. Evaluar la validez y confiabilidad de cuestionarios anteriores:
   — Adaptar un cuestionario aplicado en otro estudio.
   — Desarrollar el cuestionario propio tomando en cuenta otro(s).
5. Indicar los niveles de medición de preguntas y escalas.
6. Determinar la codificación de preguntas cerradas.
7. Elaborar la primera versión del cuestionario.
8. Consultar con expertos o personas familiarizadas con los temas investigados.
9. Ajustar la primera versión.
10. Entrenar encuestadores, si es que se necesitan (o supervisores).
11. Llevar a cabo la prueba piloto.
12. Elaborar la versión final:
    — Decidir el contexto donde se aplicara.
13. Aplicar el cuestionario:
    — Codificar las preguntas abiertas.

4.4. Aplicación del cuestionario

Existen diferentes modalidades para la aplicación de los cuestionarios. Cada una de ellas presenta una serie de ventajas e inconvenientes que se han de tener en cuenta para decir la forma más idónea para su aplicación. Distinguiremos las siguientes:

Encuesta colectiva. Esta modalidad presenta como *ventajas* las respuestas que se emiten tienen todas idénticas condiciones, se consigue un mayor índice de participación, la contaminación de unas respuestas a otras es menor y las preguntas abiertas tienen un mayor éxito. Como *desventajas*, podemos señalar, por un lado, la dificultad

de reunir a un grupo de encuestados, el área geográfica a tratar es más limitado y el retraimiento de las respuestas en diversos sectores.

Autodeterminado. En esta modalidad, el cuestionario se proporciona directamente a los respondientes, quienes lo contestan. No hay intermediarios y las respuestas las marcan ellos. Como *desventaja*, señalar que no se puede aplicar a personas que sean analfabetas, personas que tengan dificultades con la lectura o a los niños, aunque cada vez hay más cuestionarios gráficos que usan escalas sencillas.

Por entrevista personal. En este caso, es un entrevistador el que aplica el cuestionario a los respondientes. El entrevistador va haciendo las preguntas y anota las respuestas. Como *ventajas*, destacaremos la facilidad de cooperación con los sujetos y permite establecer entre ellos una relación de confianza, suele proporcionar elevadas tasas de respuesta y permite aclarar dudas en las preguntas. Como *desventajas*, surge la necesidad de tener más de un entrevistador, deben estar entrenados y conocer a fondo el cuestionario para no sesgar la respuesta, costes elevados, disponibilidad de entrevistadores y el tiempo de recogida de datos suele ser elevado.

Autoadministrado por correo postal. Consiste en el envío del cuestionario a los sujetos seleccionados con claras instrucciones sobre su cumplimentación. Como *ventajas*, señalar que los respondientes contestan directamente al cuestionario, ellos marcan las respuestas, no hay intermediarios; abarcan un área geográfica extensa y tienen un bajo coste; evita sesgos del entrevistador, y produce mayor sensación de anonimato. Como *desventajas*, podemos señalar que no hay retroalimentación inmediata y si los sujetos tienen alguna duda no se les puede aclarar en el momento, las preguntas se pueden interpretar mal, la tasa de respuesta es escasa y puede existir contagio en la respuesta en la medida en que los encuestados pueden ponerse en contacto entre sí, no existe control sobre la posible ayuda de otras personas y falta de control en el orden de las preguntas pudiéndose producir sesgos en las respuestas.

Por entrevista telefónica. Esta situación es similar a la de entrevista personal sólo que la entrevista no es cara a cara, ya que el entrevistador hace las preguntas por este medio. Entre sus *ventajas*, destacar que es más rápida que la personal; se da un mayor índice de participación; se evitan deformaciones, pues el entrevistado no conoce el sentido general del cuestionario; se necesitan pocos entrevistadores; proceso de recogida de datos corto, y permite aclarar dudas en las respuestas. Entre sus *desventajas*, señalar el hecho de no poder adjuntar información gráfica, debe ser breve, los costes pueden elevarse, sólo se puede aplicar a aquellas personas que tienen el medio, menos cooperación por parte del sujeto y escasa flexibilidad en las formas de recogida de datos.

Por correo electrónico y por internet. Estas modalidades han ido ganando terreno y ofrecen la posibilidad de interacción y asesoría. Esta modalidad tiene como *ventajas* la economía, ya que resulta muy barato; la tabulación puede hacerse de forma automática con la consiguiente rapidez en la recogida de información; registro automático de los datos, y posibilidad de referirse a/o insertar preguntas previas. Como *desventa*

jas, tan sólo puede aplicarse a personas que tienen correo electrónico y acceso a internet y pérdida del anonimato al utilizar correos personales, dificultad para usar cuestiones abiertas y riesgos derivados del medio.

5. LA ENTREVISTA

La entrevista es quizá la técnica de uso más frecuente para obtener información de la gente. pero sólo hasta hace poco ha sido utilizada con fines científicos tanto en el laboratorio como en el trabajo de campo. 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.

Desde el punto de vista de la investigación, la entrevista sirve para tres propósitos principales:

- Como un dispositivo exploratorio para ayudar a identificar variables y relaciones para sugerir hipótesis y para guiar otras fases de la investigación.
- Ser el principal instrumento de la investigación. En dicho caso, en el inventario de la entrevista se incluyen preguntas diseñadas para medir las variables de la investigación.
- Puede complementar otros métodos haciendo un seguimiento de los resultados inesperados (Kerlinger, 1986:631).

5.1. Definición, ventajas y desventajas

Una entrevista es un encuentro hablado entre dos individuos que comporta interacciones tanto verbales como no verbales. No es un encuentro entre dos personas iguales, puesto que está basado en una diferencia de roles entre los dos participantes. Aquel que se le asigna mayor responsabilidad en la conducción de la entrevista se le llama entrevistador; al otro, el entrevistado. Aunque el entrevistado puede solicitar la entrevista como consecuencia de sus propias motivaciones o necesidades y así introducir sus objetivos personales en la interacción, los objetivos de la entrevista como un sistema diádico son generalmente determinados por el entrevistador.

Entre sus características, destacamos:

- Una relación entre dos personas.
- Una vía de comunicación simbólica bidireccional, preferentemente oral.
- Unos objetivos conocidos y prefijados, al menos por el entrevistador.
- Una asignación de roles que significa un control de la situación por parte del entrevistador.

Además de estas características podemos señalar distintas ventajas y limitaciones de la entrevista como instrumento de recogida de datos.

Ventajas:

- Se trata de una relación interpersonal lo que supone un gran valor empático.
- Flexibilidad. El entrevistador puede adaptarse sobre la marcha a las necesidades del entrevistado.
- Posibilidad de observación. Además de la información verbal, el entrevistador tiene la oportunidad de observar el comportamiento del entrevistado.
- Posibilidad de registrar grandes cantidades de información.
- Posibilidad de recoger información de personas que de otra forma no hubiera sido posible.

Con respecto a las *desventajas:*

- Costo relativamente elevado. Esto se refiere a la inversión tanto en el tiempo y esfuerzo del entrevistador como en su caso del entrevistado.
- La interferencia de sesgos que puede tener variada procedencia tanto del entrevistador como del entrevistado

5.2. Tipos de entrevistas

La manera más tradicional de clasificar las entrevistas se basa en el grado de estructuración. El continuo que resulta de este tipo de estructuración queda determinado por dos conceptos polares: en un extremo, las entrevistas altamente estructuradas, y en el otro, las no estructuradas en absoluto o espontáneas. La estructuración hace referencia al grado de libertad de acción durante la interacción entre entrevistador y entrevistado. La libertad de acción se relaciona con cinco componentes:

- *Estructuración de las preguntas*, o grado de libertad que tiene el entrevistador para formular preguntas al entrevistado.
- *Estructuración de las respuestas*, o grado de libertad que tiene el entrevistado para responde a las preguntas del entrevistador.
- *Estructuración de las secuencias* de la entrevista, es decir, la libertad que tiene el entrevistador para modificar, cambiar de lugar, omitir o generar preguntas en función de la dinámica de la entrevista.
- *Estructuración del registro*, o maniobrabilidad del entrevistador para recoger y elaborar la información facilitada por el entrevistado.
- *Estructuración en la interpretación de la información*, que hace referencia a la libertad de elección de los criterios para interpretar la información.

En uno de los extremos se ubicarían las entrevistas altamente estructuradas; en ellas, tanto las preguntas como las respuestas, secuencias, formas de registro e interpretación están diseñadas de manera que el grado de libertad es nulo. El entrevistado responde a una serie de preguntas prefijadas respondiendo a algunas de las alternativas de respuesta mediante una hoja de registro donde el entrevistador interpreta según unos parámetros debidamente validados para cada tipo de población. En este tipo de entrevistas, el papel del entrevistador es prácticamente nulo, limitándose, como hemos señalado anteriormente, a interpretar según unos parámetros dados las respuestas del entrevistado y anotándolas en una hoja de registro.

En el otro extremo se encuentran las entrevistas no estructuradas en absoluto. En ellas ninguno de los cinco componentes están estructurados, esto es, la libertad para formular preguntas por el entrevistador, de responder por el entrevistado, de generar secuencias de preguntas, de registrar y de interpretar la información gozan de un grado máximo de libertad. En la realidad no es frecuente que se dé este tipo de entrevistas, ya que es difícil que el entrevistador se enfrente al entrevistado sin ningún tipo de guión, aunque sólo sea implícito (Fonfría-Bonavia, en Quintanilla, 1992).

Sin llegar a estos extremos en el continuo, podemos hablar de forma general de entrevistas estructuradas, semiestructuradas y no estucturadas.

En las entrevistas estructuradas, también llamadas normalizadas o dirigidas, el entrevistador actúa bajo un esquema establecido de interacción que incorpora preguntas prefijadas de antemano y a las que cualquier entrevistado debe responder de manera más o menos cerrada sin demasiada libertad de respuesta. Los objetivos, los contenidos y las técnicas de actuar están claramente determinados y previstos de antemano. El entrevistador tiene un papel directivo, siendo el agente que controla la entrevista. Por su parte, el entrevistado se limita a responder a las preguntas formuladas sin apenas lugar para otras incursiones. Son propias de situaciones formalizadas, solemnes y cargadas de artificialidad. Tanto el entrevistado como el entrevistador tienen muy bien delimitado sus papeles y deben circunscribirse a ellos, lo que provoca rigidez, formalismo y una cierta restricción a la hora de formular las preguntas y elaborar respuestas.

Con respecto a la secuenciación de este tipo de entrevistas, el entrevistador carece de libertad para acomodar sus preguntas a la información verbal y/o no verbal que expone el entrevistado. Esto repercute en la falta de profundidad de algunos temas que pueden ser importantes para la consecución de los objetivos de la investigación al no poder generar, omitir o cambiar preguntas para adentrarse en los tópicos que puedan parecer de interes.

El registro e interpretación de la información son aspectos diferenciadores de los tipos de entrevistas. Una estructuración de los datos ayudará a la cuantificación de los mismos y a la toma de decisiones. Con la estructuración de la información se otorgan unas reglas objetivas que obligan al entrevistador a tomar decisiones a partir de los resultados alcanzados en función de unas variables determinadas.

Las entrevistas no estructuradas, no dirigidas o no normalizadas son aquellas en las que el entrevistador dirige la entrevista con un esquema altamente flexible en la formulación de las preguntas y otorga al entrevistado una gran libertad de respuesta. Este tipo de entrevista se caracteriza por no tener determinados los objetivos ni los contenidos a tratar. La finalidad esencial se centra en la propia realización, en la propia entrevista. Debe realizarse en las condiciones más naturales posible huyendo de la artificialidad.

El entrevistador es lo menos directivo posible, acercando la entrevista a un diálogo simétrico. Son especialmente indicadas para establecer primeros contactos indicar puntos de vista, alcanzar primeros acuerdos, etc.

Con respecto a la secuenciación en las entrevistas no estructuradas, el entrevistador puede variar en cualquier momento su estrategia de conversación para introducirse en los temas que parezcan más convenientes, obteniendo así una información de mayor calidad.

Referente al registro e interpretación de la información en la entrevista no estructurada, se carece de la uniformidad de evaluación, ya que la no estructuración y la búsqueda de calidad complican su cuantificación.

Tal y como hemos dicho, los tipos de entrevistas se mueven en un continuo donde existen grados de estructuración bajo los cuales el entrevistador puede moverse para conseguir sus objetivos de la mejor manera posible, hablamos ahora de las entrevistas semiestructuradas o mixtas. Este tipo de entrevistas comparte las ventajas e inconvenientes de ambos en mayor y menor medida según se acerquen a un lado u otro. En función de los objetivos a alcanzar en la investigación, se ha de diseñar la entrevista que mejor se adecúe para su consecución. Bajo la definición de entrevista como un proceso de interrelación, se hace necesaria una determinada retroalimentación entre entrevistado y entrevistador en la que se siga un cierto grado de libertad en el comportamiento de ambos de la misma forma que algunas variables son difíciles de estudiar únicamente con una lista de preguntas prefijadas, siendo preciso recurrir a un grado de interacción mayor.

5.3. Reactivos o preguntas

Al igual que en los cuestionarios, en la entrevista podemos encontrar tres tipos de reactivos o preguntas: de *alternancia fija o cerradas, abiertas* y *de escala*.

Alternancia fija. Como su nombre indica, este tipo de preguntas ofrece al entrevistado una opción entre dos o más alternativas. El tipo más común es el dicotómico: plantea preguntas que pueden responderse como *sí* o *no*, *de acuerdo* o *en desacuerdo* y otro tipo de respuesta de dos opciones. Con frecuencia, se añade una tercera alternativa fija: *no sé*. Entre las *ventajas* de este tipo de preguntas es lograr una mayor uniformidad en la medición y, por tanto, mayor confiabilidad, forzar al entrevistado a responder de una forma que se ajuste a las categorías previamente

establecidas y ser fáciles de codificar. Entre sus *desventajas*, está la superficialidad, puede irritar al entrevistado al no encontrar ninguna alternativa adecuada para él.

Reactivos abiertos. Este tipo de preguntas son aquellas que brindan un marco de referencia para las respuestas de los entrevistados, pero poniendo un mínimo de restricción a las respuestas y a su expresión. Aunque su contenido está determinado por el problema de investigación, no imponen ninguna otra restricción sobre el contenido ni sobre la forma de respuesta del entrevistado. Este tipo de pregunta es flexible, tienen la posibilidad de profundizar y le permiten al entrevistador aclarar malos entendidos. Un tipo de pregunta abierta es la pregunta embudo; recordemos que este tipo de pregunta se inicia con una pregunta general hasta llegar progresivamente a preguntas más específicas.

Reactivos de escala. Una escala es un conjunto de reactivos verbales a cada uno de los cuales un individuo responde expresando grados de acuerdo o desacuerdo o algún otro modo de respuesta. Los reactivos de escala tienen alternativas fijas y colocan al individuo encuestado en algún punto de la escala.

Independientemente del tipo de reactivo o preguntas que usemos en la entrevista, hay una serie de criterios necesarios para la buena redacción de las mismas. Estos criterios son:

a) Ha de estar relacionada con el problema y los objetivos de la investigación. Esto significa que el propósito de la pregunta es generar información para probar las hipótesis de la investigación.

b) Se ha de elegir el tipo de pregunta adecuado a la investigación. Alguna información puede obtenerse mejor con preguntas abiertas, otras cerradas o escalas; la buena elección de las mismas beneficiará la recogida de información y posterior interpretación de resultados.

c) Han de ser claras y sin ambigüedades. Una pregunta ambigua es aquella que permite o invita a interpretaciones alternativas de las cuales resultan respuestas diferentes; se ha de evitar la ambigüedad en las preguntas de la entrevista.

d) No han de ser conducentes, es decir, no han de dirigir ni sugerir la respuesta del entrevistado.

e) Se ha de tener cuidado con las preguntas que demanden información que el entrevistado no posee. Antes de hacer al entrevistado una pregunta sobre un tema, por ejemplo, CSIC, hay que asegurarse si el entrevistado sabe lo que es el CSIC y después preguntar sobre él.

f) Las preguntas no han de ser comprometidas para el entrevistado; si debe tratarse un tema personal ha de hacerse de forma sigilosa garantizando la confidencialidad.

g) Evitar preguntas con respuestas estereotipadas, es decir, preguntas cargadas de aceptación social.

5.4. Fases de la entrevista

Las fases de la entrevista son:

1. Preparación o planificación
2. Ejecución.
3. Evaluación.
4. Control.

Preparación o planificación. Esta primera fase o de planificación consiste en una elaboración por parte del entrevistador del esquema que seguirá en la interrelación. Este esquema hace referencia a la estructuración de la entrevista en sus cinco dimensiones, a los objetivos que se pretenden y a la forma de conseguirlos.

La *selección de los entrevistados* es un proceso importante, unas veces puede ser una persona determinada porque interese para la investigación que así sea o puede ser un grupo de personas que muestren características semejantes dentro de la población que reúne las características que queremos estudiar.

El entrevistador debe *elegir el momento y el lugar* para llevar a cabo la entrevista. Tanto si se trata de una persona determinada como de un grupo, se ha de concertar previamente la misma y buscar un lugar donde los entrevistados o el entrevistado se encuentren cómodos con buenas condiciones ambientales.

Ejecución y registro. En este segundo momento es donde se realiza la entrevista propiamente dicha, es decir, es la situación en la cual se da un contacto directo entre entrevistador y entrevistado. Esta fase es lo suficientemente amplia y compleja que hace que no sea homogénea. Podemos distinguir tres momentos:

A un primer momento le llamaremos *etapa de comienzo o toma de contacto*. En esta fase el entrevistador debe crear un ambiente de confianza y lograr una aceptación recíproca de la interacción. El entrevistador debe presentarse a sí mismo, así como indicar los motivos de la misma, y seguir con un proceso de diálogo donde se den preguntas sencillas relacionadas o no con el objetivo de la entrevista. El entrevistador debe reforzar cualquier conducta del individuo que esté relacionada con la disminución de la tensión y la consecución de un informe adecuado para el posterior desarrollo de la interacción. Para ello, el entrevistador debe acercarse al entrevistado de forma amistosa procurando que se sienta seguro y dispuesto a hablar.

Al segundo momento se le denomina *cuerpo de la entrevista*. Una vez que se ha logrado una predisposición del entrevistado para participar en el proceso, tiene lugar la fase de la entrevista donde el intercambio de información se vuelve más intenso. En ella

se produce su recogida y su contrastación de acuerdo con el guión previo del entrevistador y el grado de estructuración previsto. Es la etapa de mayor duración y de mayor implicación por ambas partes, dependiendo su éxito de la capacidad de la entrevista para abordar de manera satisfactoria los puntos a tratar y de la habilidad del entrevistador para conseguir y registrar la información proporcionada por el entrevistado. Para conseguir este clima, es importante seguir algunas orientaciones como no dar la impresión a la hora de hacer las preguntas que se trata de un interrogatorio o examen, para lo cual se evitarán preguntas que puedan denotar sorpresa, duda o crítica; mostrar interés por las respuestas emitidas; las preguntas han de sucederse con cierta rapidez con el fin de que las respuestas sean lo más espontáneas posibles; el entrevistador no deberá exponer su opinión personal en ningún aspecto.

El tercer momento es la *terminación*. Cuando el esquema de la entrevista ha sido cumplimentado y la información recogida y contrastada, no queda más que concluir el proceso. En esta última fase, el entrevistador debe dar la oportunidad al entrevistado para aclarar los puntos que no haya comprendido o las dudas que hayan podido surgir. La finalización de la entrevista debe producirse en un clima de cordialidad y agradecimiento por las informaciones recogidas dejando abierta la posibilidad de establecer otras colaboraciones.

Con respecto al registro, deberá realizarse de la manera indicada al inicio de la misma. Existen diferentes modalidades que deberán ser valoradas en función de las preferencias y posibilidades que ofrezca el entrevistado. Si bien es aconsejable anotar cada respuesta en el momento que se producen o bien grabarlas en algún soporte (vídeo, casete), hay ocasiones en las que el entrevistado podrá mostrar desconfianza con el empleo de estos medios, por lo que se aplazará el registro de datos para el momento posterior al desarrollo de la entrevista sin que medie demasiado tiempo para que podamos reproducir del modo más exacto las respuestas emitidas (De Lara Guijarro-Ballesteros Velázquez, 2001:323).

Evaluación. Éste es el último momento de la entrevista. El entrevistador entra en un proceso de interpretación y evaluación de la información que dará lugar a un informe donde se reflejaran las respuestas recogidas en el transcurso de la misma, los incidentes surgidos, así como cuanta información pueda ayudar para el posterior análisis e interpretación de la información.

Otro punto importante es el **control** de la entrevista. La entrevista es un instrumento de recogida de información, debiendo ser ésta válida y fiable. Para conseguir esto, la entrevista debe estar sometida a pruebas de fiabilidad y validez como cualquier otro instrumento de medida. Las técnicas que podemos utilizar pueden ser los procedimientos mencionados al estudiar la fiabilidad y la validez de los instrumentos de medida, como el juicio de expertos y el control de los ítem o preguntas que se mencionan en la entrevista.

6. PRUEBAS E INVENTARIOS ESTANDARIZADOS

En la actualidad, existe una gran variedad de pruebas estandarizadas para medir gran número de variables: la motivación, el rendimiento, la personalidad, la inteligencia, habilidades, valores, intereses, pruebas clínicas, pruebas sociales, competencias profesionales, etc. Todas estas variables se miden con estas pruebas que tienen su propio procedimiento de aplicación, codificación e interpretación; entre estas pruebas estandarizadas se encuentran los test.

6.1. Definición y características

Conbach (1972) lo define como «un procedimiento sistemático para observar la conducta y describirla con la ayuda de escalas numéricas o categorías establecidas».

García Marcos (1983) lo define como «métodos estandarizados de recogida de información que es posible, en la mayoría de los casos, cuantificar y, por tanto, comparar los resultados con grupos normativos de referencia».

Grizb (1981) lo define como «técnicas que miden constructos teóricos definidos operacionalmente a través de los distintos ítem que lo integran».

Sea cual sea su definición, tienen como *características* las siguientes:

- Constituyen una forma de medición indirecta del rasgo o característica estudiada en cuanto que no tienen en cuenta las respuestas por sí mismas, sino por su significado en relación con la conducta, actitud, opinión o personalidad del individuo.
- Permite una descripción cuantitativa y controlable del comportamiento de un individuo ante una situación específica tomando como referencia el comportamiento de los individuos de un grupo definido colocado en la misma situación.
- Tipificación de la medida.
- Objetividad. Esto implica independencia del juicio de la persona que lo aplica.
- Están asociados al escalonamiento de sujetos.
- Utilizan condiciones estandarizadas para la recogida de datos, es decir, cada test se aplica a los sujetos bajo las mismas normas o condiciones en el momento de recoger los datos.
- Utilizan normas para medir y evaluar las características psicológicas asociadas al test.
- Poseen una alta fundamentación científica basada en teorías.
- Deben poseer los requisitos de fiabilidad y validez para garantizar la adecuación de su uso
- Permiten la predicción o inferencia.

6.2. Clasificación

La clasificación de los test puede ser muy amplia y variada dependiendo del criterio que se use; los criterios pueden ser amplios o específicos. Pueden ser clasificados por las posibilidades de aplicación, por el tipo de material a utilizar, según exista o no tipificación, según el propósito de la medida, etc. No es el interés de este apartado presentar una amplia y extensa clasificación de los test, ya que lo que se pretende es una visión general de los mismos.

Los tres tipos principales de diseños de test psicométricos que se han configurado durante el siglo pasado son los siguientes:

- Diseños de test de norma de grupo.
- Diseños de test referidos a criterios:
  - — Referidos a dominio.
  - — Referidos al objetivo.
- Los diseños de test adaptativos.

Diseños de test de norma de grupo. Estos test tienen como objetivo medir características o propiedades psicológicas de cada sujeto perteneciente a un grupo normativo en relación con los restantes sujetos de dicho grupo normativo.

Para ello, se comparan las respuestas al test de cada sujeto, valoradas numéricamente, con medidas de tendencia central (media y/o mediana) y con medidas de desviación (puntuaciones *z*, puntuaciones *z* derivadas, centiles, etc.) obtenidas por el grupo normativo correspondiente.

En consecuencia en este tipo de test la puntuación o rango obtenido(a) no significa nada; para dar su significado hay que convertirlo en una puntuación o rango estándar de norma estadística utilizando para ello medidas estadísticas de tendencia central o desviación.

Los test referidos a criterio. Se diferencian de los de norma de grupo en que no pretenden medir diferencialmente características psicológicas de los sujetos de un grupo normativo a partir de su comparación estadística, sino que miden y valoran a cada sujeto individualmente en relación con el nivel alcanzado en un determinado dominio previamente delimitado, definido y escalado.

En este tipo de test, al contrario que el anterior, las puntuaciones o rasgos asociados a la escala tienen un significado por sí mismo, éste puede ser numérico o cualitativo o sólo numérico. Uno u otro significado depende de que el dominio asociado al test esté delimitado, definido y escalonado o no jerárquicamente.

Dentro de los test referidos a criterio, hay dos variedades: *referidos a dominio* y *referidos al objetivo*.

Referidos a dominio, tienen como finalidad medir y evaluar el estado de los cambios individuales en un determinado dominio con fines de diagnóstico psicológico, pedagógico, académico, etc.

Referidos al objetivo, tienen una finalidad prioritariamente selectiva al establecer uno o más puntos de corte en la escala asociada a un determinado domino. Este punto o estos puntos de corte se utilizan como criterio para tomar decisiones en cada caso. Un ejemplo, el punto de corte de una prueba selectiva o un examen, ya que estos puntos de corte determinan o no la selección para un puesto de trabajo o distintas calificaciones de un examen.

Test adaptativos. Este tipo de test puede estar asociado a los test de norma de grupo o a los referidos a criterio. En ambos casos, la administración del test es individual y adaptan el nivel de dificultad de sus ítem al nivel de capacidad del sujeto durante el proceso de aplicación del test. Utilizan un punto de partida diferencial para sujetos diferentes, estableciendo reglas para seleccionar los ítem siguientes en función de las respuestas que hayan dado en las precedentes. La aplicación del test concluye para cada sujeto cuando se ha estabilizado un determinado nivel o tipo de respuesta por parte del sujeto (Losada-López-Feal, 2003).

6.3. Elaboracion

En este apartado vamos a indicar los pasos generales a seguir para la elaboracion de un test psicométrico de norma de grupo.

1. Título del test. A través del título se identifica el test que se pretende construir. Esta identificación viene determinada, fundamentalmente, por el tipo de conductas, contenido o constructo que pretendemos medir a través de él.

2. Antecedentes. En este apartado se hace referencia, si procediera, a otros test precedentes o contemporáneos que tienen una relación más o menos directa con el test que se pretende construir.

3. Objetivos. Los objetivos del test vienen determinados conjuntamente por las respuestas que se den a las preguntas *¿qué* vamos a medir realmente a través del test que nos proponemos construir? y *¿cuál* es la finalidad para la que construimos dicho test?

4. Marco teórico y diseño asociado al test. Es el momento de señalar el marco teórico o teoría que sustenta al test.

5. Metodología de recogida de datos asociada a la elaboración o adaptación del test con las siguientes subetapas:

   - Preselección de ítem.

   - Preselección de sujetos.

- Estandarización de las condiciones internas y externas asociadas al test.
- Formato de los ítem del test.
- Especificación de los criterios de puntuación y valoración de las respuestas de cada ítem y a su conjunto.
- Aplicaciones del test a los sujetos preseleccionados.
- Puntuación y valoración de las respuestas.
- Tabulación de las respuestas y otras variables.

6. Metodología de tratamiento de datos asociada a la selección del test como instrumento de medida: alternativas de análisis de datos. La finalidad de esta etapa es la de seleccionar aquellos ítem que constituyen los indicadores más adecuados para representar los contenidos y comportamientos indicados en la parte teórica.
7. Metodología de tratamiento de datos asociada a la objetivización del test seleccionado: la estimación de coeficientes de fiabilidad y la estimación o validez del test como instrumento de medida. Lo que se pretende en esta etapa es comprobar en qué grado son precisas y consistentes las medidas del test estimando para ello coeficientes de fiabilidad, validez y, en su caso, sus complementarios errores de medida al azar o aleatorios.
8. Metodología de tratamiento de datos asociada a la obtención de puntuaciones de rangos estandarizados: estimación de baremos obtenidos por el grupo normativo. Una vez que el test de norma de grupo ha superado todos los procesos teórico-metodológicos, llega el momento de transformar las puntuaciones directas o ponderadas obtenidas en puntuaciones comparables que permitan luego clasificar a los sujetos por niveles a partir de una norma en este caso estadística (Losada López-Feal, 2003).

7. LA OBSERVACIÓN

La observación es una técnica útil para el investigador, consiste en un conjunto de registros de incidentes de comportamiento que tienen lugar en el curso normal de los acontecimientos y que son destacados como significativos para describir modelos de desarrollo. En la vida diaria, todos observan los actos de todos, se observa a otras personas, se les escucha hablar, se infiere lo que otros quieren decir, pero la observación científica es otra cosa. El científico busca observaciones confiables, objetivas, a partir de las cuales puedan realizar inferencias válidas.

Existe mucha controversia sobre la observación: por un lado, están quienes opinan que la observación debe responder a unos criterios regulados rigurosamente obteniendo como críticas que es un proceso rígido y artificial, y por otro, quienes opinan que la observación se debe realizar en un ambiente natural en el que los observadores deben

estar inmersos en situaciones realistas. A la primera, la podemos denominar observación sistemática, señalando distintos grados de sistematización, siendo más utilizada en el enfoque cuantitativo, y a la segunda, observación participante, diferenciando igualmente distintos grados de participación, usándose más en un enfoque cualitativo. Sin duda, la observación, ya sea bajo un enfoque u otro, es un instrumento valiosísimo en la recogida de datos y serán las características de la investigación las que determinarán el uso de un tipo u otro.

En este capítulo trataremos la observación desde un enfoque cuantitativo, dejando el enfoque cualitativo para otro tema posterior.

Desde un enfoque cuantitativo, definimos la observación como un registro sistemático, válido y confiable de comportamiento o conducta manifiestos. En la observación sistemática, los eventos son seleccionados, registrados, codificados en unidades significativas e interpretados por no participantes (Dane, 1990:151). Uno de los aspectos más importantes consiste en decidir cuáles son los eventos de interés y, más particularmente, cómo han de ser seleccionados. El interés principal de la observación sistemática gira, pues, en torno a los eventos cuya categorización, propuesta por Weick (1968), es la siguiente: eventos no verbales, espaciales, extralingüísticos y lingüísticos. Un segundo tema de interés es el relativo a su selección, donde están implicados el tiempo o el evento en sí mismo. Las técnicas más importantes que se han propuesto pueden ser muestreo de tiempo continuo, muestreo de tiempo puntual o instantáneo, muestreo de intervalos de tiempo y muestreo de eventos (Anguera, 1998:39).

La observación sistematizada se lleva a cabo mediante un procedimiento planificado previamente en el que queda explícito tanto el objetivo de la medición como la forma de registro de los datos de acuerdo con una norma establecida. El investigador posee unos conocimientos previos acerca de la realidad objeto de estudio que son los que permiten establecer la planificación del proceso de observación.

7.1. Pasos para construir un sistema de observación

Los pasos para construir un sistema de observación son los siguientes:

1. Definir con precisión el universo de aspectos, eventos o conductas a observar. Por ejemplo, supongámos que queremos investigar sobre el grado de delincuencia de algunas escuelas de secundaria de sectores marginales en una gran ciudad. Un evento a observar sería las conductas de comportamiento de los alumnos en el recreo. Otro sería el comportamiento por los pasillos y en clase en las horas de intercambio de profesorado.

2. Extraer una muestra representativa de aspectos, eventos o conductas a observar.

3. Establecer y definir las unidades de observación. Siguiendo nuestro ejemplo, una unidad de observación podría ser cada vez que se muestra una conducta agresiva.

4. Establecer y definir las categorías y subcategorías de observación. Hernández Sampieri (2003) señala las siguientes categorías:

   a) Distancia física entre el observado y el observador.

   b) Movimientos corporales que denotan tensión, relajación u otros por parte del observador.

   c) Conducta visual del sujeto:

      — Dirigida hacia el observador.

      — Dirigida hacia otra parte.

   d) Conducta verbal:

      — Frases u oraciones completas.

      — Frases u oraciones dicotómicas.

5. Seleccionar a los observadores. Los observadores son las personas que codifican la conducta y deben conocer las variables a tratar, categorías y subcategorías.

6. Elegir el medido de observación. La conducta y sus manifestaciones pueden codificarse por distintos medios: observarse directamente y codificarse; otras veces se codifican a posteriori, etc.

7. Elaborar las hojas de codificación. En ellas han de aparecer tanto las categorías como subcategorías a tratar.

8. Calcular la confiabilidad de los observadores. Existen varias fórmulas, una de las cuales puede ser el grado de acuerdo interobservadores, la confiabilidad individual, la confiabilidad por parejas, etc.

9. Proporcionar entrenamiento a los codificadores.

10. Llevar a cabo la codificación.

11. Vaciar los datos de las hojas de codificación y obtener totales para cada categoría.

12. Realizar los análisis apropiados.

7.2. Registro de datos

La forma de registrar los datos en la observación puede ser variada. Ya hemos dicho que la observación puede tener distintos enfoques; asimismo, la forma de registrar esos datos también será distinta según el enfoque del que se trate. No hay un acuerdo unánime a la hora de catalogar las formas de registro ni todos los autores barajan idénticos criterios. El criterio que proponemos a continuación se basará en el grado de la es-

tructuración de la observación. En la observación altamente estructurada utilizaríamos sistemas categoriales altamente estructurados, como Bales y Flandes, propios del enfoque cuantitativo; en la estructurada utilizaríamos como registro datos, también propios del enfoque cuantitativo, las *listas de control*, los *sistemas de signos* y las *escalas de estimación*, y en la observación no estructurada, el diario, las notas de campo y el registro de incidentes críticos, propios del enfoque cualitativo.

En este caso, nos centraremos en las listas de control, los sistemas de signos y las escalas de estimación. No es motivo de este tema hacer un estudio exhaustivo de cada una de estas formas de registro, pero sí dar unas pequeñas indicaciones para que el lector conozca algo de ellas.

Listas de control. Consisten en relaciones estructuradas de características, habilidades, cualidades sociales, rasgos de conducta, secuencia de acciones, etc. Proporcionan información sobre la presencia o ausencia de estos rasgos sin determinar la intensidad o frecuencia del mismo. Las preguntas se presentan con una estructura de respuesta de:

❑ sí ❑ no

Al igual que en el resto de los instrumentos de medida, a la hora de elaborar una lista de control se ha de tener en cuenta el objetivo de la investigación, los rasgos característicos o acciones, la definición operativa de los rasgos y la estructuración de las respuestas.

A partir de la información obtenida de las listas de control, se pueden conocer aquellos aspectos que resultan deficitarios para posteriormente idear un plan de actuación para su mejora.

Sistemas de signos. Es un tipo de sistema de observación centrado en el examen de conductas específicas que son registradas por el observador sin emitir valoración alguna sobre ellas. El objetivo es muestrear numerosas porciones de un evento natural sin que exista ninguna referencia a la importancia en una dimensión. Lo que se registra es la presencia o ausencia de ciertas conductas y si es pertinente su frecuencia de aparición. Según Anguera (De Lara Guijarro, 2001), la estrategia a seguir en el sistema de signos es dividir el tiempo de observación en períodos breves de tres a cinco minutos y registrar cada signo la primera vez que ocurre dentro del intervalo. La frecuencia de cada intervalo será 0 o 1, según se haya dado o no. El recuento total no mostrará el número de veces que ha aparecido la conducta, sino el número de intervalos en la cual se ha mostrado.

Escalas de estimación. Es un instrumento de medición compuesto por un conjunto de símbolos o valores numéricos construido de tal manera que los símbolos o valores numéricos puedan ser asignados por una regla a los individuos a quienes se les aplica la escala y donde la asignación indica si el individuo posee lo que se supone que mide la escala.

Las escalas de estimación pueden ser:

De categorías, en las que se presenta a los observadores o jueces varias categorías, donde ellos eligen la que mejor caracteriza el comportamiento o característica del objeto que se estudia; esas categorías pueden ser siempre, frecuentemente, raras veces, nunca.

De números, donde la calificación se hace por números: 4, 3, 2, 1, 0, con cada reactivo o a la inversa.

Gráfica donde se combinan líneas o barras con frases descriptivas

muy fácil	fácil	poco fácil	nada fácil

Para la elaboración de una escala de estimación, al igual que en el resto de los instrumentos de medida, se ha de tener en cuenta los objetivos específicos de forma clara que queremos valorar, la selección de los rasgos y la graduación de la escala y el número de categorías.

BIBLIOGRAFÍA

ANGUERA, M. T. (1989): *Manual de prácticas de la observación en las ciencias humanas*. Madrid: Cátedra.

ANGUERA, M. T.; ARNAU, J.; ATO, M., y otros (1998): *Métodos de investigación en psicología*. Madrid: Síntesis.

BALCELS I JUNYET, J. (1994): *La investigación social. Introducción a los métodos y las técnicas*. Barcelona: Promociones y Publicaciones Universitarias.

BARBERO, M. I. (1999): *Psicometría, II. Métodos de elaboración de escalas*. Madrid: UNED.

BOHRNSTEDT, G. W. (1976): «Evaluación de la confiabilidad y validez en la medición de actitudes», en SUMMERS, G. F. (comp.), *Medición de actitudes*. México: Trillas, pp. 103-127.

BRIOMES, G.: *Métodos y técnicas de investigación para las ciencias sociales*. México: Trillas.

BUENDÍA, L. (1997): «La investigación por encuesta», en COLÁS, P.; BUENDÍA, L., y HERNÁNDEZ PINA, F., *Métodos de investigación en psicopedagogía*. Madrid: McGraw-Hill, pp. 120-157.

COHEN, J. (1960): «A coefficiente of agreement for nominal scales», *Educational and Psychological Measurement*, 20, 37-46.

CONBACH, L. J. (1972): *Fundamentos de la exploración psicológica*. Madrid: Biblioteca Nueva.

DANE, F. C. (1990): *Research methods*. Pacific Grove, CA: Brooks/Cole.

Reyes Lagunes, I., García y Barragán, L. F., . *La Psicología Social en México* (Vol. 12). México/ AMEPSO, UNAM Facultad de Psicología , División de Ciencias Sociales de la Universidad de Sonora. Capítulo 87 Pp 623-630

AMEPSO

LA PSICOLOGÍA SOCIAL EN MÉXICO

VOLUMEN XIII

Comité Editorial

Sofía Rivera Aragón

Rolando Díaz Loving

Isabel Reyes Lagunes

Rozzana Sánchez Aragón

Luz María Cruz Martínez

Coordinación Editorial

Sofía Rivera Aragón

Luz María Cruz Martínez

Primera edición: 2010

ISBN 968-5411-13-1

Este libro se publicó con apoyo del Programa
Integral de Fortalecimiento Institucional
P/PIFI 2009-26MSU0015Z-09

AMEPSO

La Psicología Social en México

VOLUMEN XIII

explicada, mediante la elaboración de reactivos a partir de respuestas de las personas a preguntas abiertas u otras técnicas exploratorias.

Referencias

Bugaighis, M., Schumm, W., Bollman, S., & Jurich, A. (1983). Locus of control and marital satisfaction. *The Journal of Psychology, 114*, 275 – 279.

La Rosa, J. (1986). *Escalas de locus de control y autoconcepto.* Tesis de doctorado no publicada. Facultad de Psicología. UNAM.

Leone, C. & Burns, J. (2000). The measurement of locus of control: Assessing more than meets the eye?. *The Journal of Psychology, 134* (1), 63-76.

Doherty, W. (1983). Locus of control differences and marital dissatisfaction. Journal of Marriage and the Family, *43*, 369 – 377.

Doherty, W., & Ryder, R. (1979). Locus of control, interpersonal trust and assertive behavior among newlyweds. *Journal of Personality and Social Psychology, 37*, 2212 – 2220.

Marks, L., (1998). Deconstructing locus of control: Implications for practitioners. *Journal of Consueling and Development, 76* (3), 251 – 260.

Miller, P., Lefcourt, H., Holmes, J., Ware, E., & Saleh, W. (1986). Marital locus of control and marital problem solving. *Journal of Personality and Social Psychology, 51* (1) , 161-169.

Myers, S., & Booth., (1999). Marital strains and marital quality: the role of high and low locus of control. *Journal of Marriage and the Family, 61* (2), 423 – 436.

Reyes, I., (2007, julio). Escala Mexicana de Locus de Control. Simposio efectuado en el VI Congres Iberoamericano de Evaluación Psicológica, D.F., México.

Robinson, J., Shaver, P., (Eds), & Wrightsman, L. (1991). *Measures of Personality and Social Psycholog Attitudes: Volume 1: Measures of Social Psychological Attitudes.* USA: Academic Press.

PROCEDIMIENTO DE VALIDACIÓN PSICOMÉTRICA CULTURALMENTE RELEVANTE: UN EJEMPLO

ISABEL REYES LAGUNES[1] Y LUIS FELIPE GARCÍA Y BARRAGÁN

Universidad Nacional Autónoma de México

Es por todos conocidos que para que una propuesta de conocimiento se convierta en ley universal y de paso al conocimiento científico es que éste debe poder ser replicado. La Psicología no podría haber sido reconocida como ciencia sin contar con esta norma. Ernst Heinrich Weber y Gustav Theodor Fechner establecen, a principios del siglo XIX, la relación cuantitativa entre la magnitud de un estímulo físico y como éste es percibido abren claramente el papel de la medición en la Psicología que es reconocida como ciencia con base en la labor realizada por Wilhelm Maxiliam Wundt y sus estudiantes en Liepzig en 1879. A partir de este momento y obviamente con la influencia de aportaciones de Charles Darwin, Francis Galton y Binet, entre otros, que llevan a James McKeen Cattell, en 1890, a identificarlos como Tests Mentales. Interesantemente, desde 1940, Raymond Bernard Cattell apunta a la necesidad de desarrollar instrumentos de medición 'libres de cultura' y/o 'culturalmente justas', lo cual lleva necesariamente a una clara definición del concepto cultura para su operacionalización y medición. En nuestro país destaca la labor de Rogelio Díaz Guerrero, pionero en esta tarea, cuando la define como: El sistema de premisas socioculturales interrelacionadas que norman y gobiernan los sentimientos, las ideas, la jerarquización de las relaciones interpersonales y estipula, los papeles sociales que hay que llenar, las reglas de interacción de los individuos en tales papeles, los dónde, cuándo, con quién y cómo desempeñarlos (Díaz-Guerrero, 1961). Tal y como afirma, Clyde Kluckhon: Cultura es a Sociedad lo que memoria es al individuo. Desafortunadamente, a pesar del generalizado acuerdo con la importancia de la cultura en la medición, los esfuerzos para su vigilancia han sido pocos.

Con base en lo anterior y tomando en consideración que existen un gran número de pruebas psicométricamente válidas pero desarrolladas en otros entornos culturales que presentamos, a través de un ejemplo, propuesta de metodología para su adaptación a otra cultura. Para ilustrar la propuesta metodológica-estadística de este trabajo, se tomará

[1] Correo electrónico: reyesisa@yahoo.com.mx

como ejemplo el proceso de validación de una escala de Persistencia y Perseverancia, desarrollada a partir de la escala presentada en el proyecto "International Personality Item Pool" (2008) basada en la escala de *Values in Action* de Peterson y Seligman (2004) que originalmente cuenta con una consistencia interna de Chronbach de .81.

Propuesta metodológica-estadística para adaptación y validación psicométrica

A continuación presentamos cada uno de los pasos propuestos

1. Traducción vigilando
   a) Equivalencia del lenguaje
   b) Equivalencia cultural
      Forma
      Transformación (transformar radicalmente el reactivo para llegar al contenido)
      Radical

A continuación se realizó

1. Validación por jueces
2. Re-traducción a idioma original(una persona que no conozca el instrumento)
3. Adecuación
4. Piloteo
5. Revisión y re-adecuación
6. Aplicación a población meta
7. Validación psicométrica a través de los siguientes análisis:
   a) Frecuencias incluyendo sesgo para selección de reactivos dependiendo del objetivo del instrumento (conductas típicas o normales).
   b) Discriminación de reactivos para grupos extremos con base en el cuál se eliminarán los que no cumplen con el requisito.
   c) Confiabilidad interna
   d) Validez

Método

Participantes

La muestra estuvo conformada por 356 participantes, 188 mujeres (52.8%) y 168 hombres, en un rango de edades entre los 15 y 70 años, que en conjunto reportan una media de edad de 24 años (DE = 10). 226 (63.5%) participantes reportan ser estudiantes, mientras que el resto, 130, declara dedicarse al hogar o a una actividad remunerada

Instrumento

Siguiendo el orden establecido en el procedimiento se inició por hacer la traducción donde palabras tales como *love* (*Love to think…, Love to read…*), *pleasure* (*Put work above pleasure*), y expresiones tales como: *Have an eye for detail* no pueden ni deben traducirse si no adaptarse a términos utilizados en el medio y potenciales respondientes del instrumento tomando en consideración qué se intenta medir. Frecuentemente esta tarea requiere presentar a jueces más de una opción de verbalización hasta lograr acuerdo mayoritario entre ellos. Una vez logrado el consenso se re-tradujo al idioma original, se revisó la pertinencia del lenguaje y se realizó el estudio piloto de la 1ª. versión para realizar ajustes de ser necesario.

Con base en lo anterior, se desarrolló un instrumento de autoaplicación con cinco opciones de respuesta tipo Likert, integrado por 55 reactivos que junto con la *Persistencia y perseverancia* evaluaron otros cinco constructos teóricos relacionados. Estos reactivos e complementaron con una sección de datos generales (edad, sexo, ocupación, etc.).

Procedimiento

El encuestador contactó a los participantes en sus centros de estudio, su hogar o trabajo y estuvo presente durante la respuesta del instrumento para resolver las dudas que pudieron surgir en el proceso. En todos los casos se remarcó el hecho de que los datos obtenidos serían tratados con fines estadísticos y en ningún momento se le pidió al participante su nombre o cualquier dato con el que pudiera identificársele. Una vez que se obtuvieron los datos, se generó una base de datos y se aplicó el proceso de validación estadística mencionado en la parte respectiva.

Resultados

En primer lugar se realizó un análisis de frecuencia para cada uno de los reactivos de la escala de Persistencia y Perseverancia, solicitando Media, desviación estándar y Sesgo, con el fin de a) verificar que esté bien capturado (ningún valor fuera de rango), y que todas las opciones de respuesta hayan sido atractivas, es decir existe frecuencia en cada una de ellas, b) la direccionalidad de los reactivos con base en lo que se intenta medir, ver Tabla 1 y c) dependiendo del objeto de estudio (conductas típicas o conductas que poseen una distribución normal) será el sesgo que estaremos esperando. Si es el primero de los casos (típicas) el sesgo debe ser menor a -.5 y mayor a +.5. Si por el contrario se busca la curva normal, los valores a buscar deberán encontrarse entre -.5 y .5.

Tabla 1. Análisis de frecuencias de dos reactivos ejemplos de la escala de Persistencia y Perseverancia

	Desacuerdo	1	2	3	4	5	Acuerdo
No abandono una tarea hasta que la he concluido		37	66	110	113	119	
Nunca termino lo que comienzo		165	108	65	56	51	

Como se puede observar en esta Tabla, el segundo de los reactivos es contrario al objetivo del instrumento, por lo tanto, su calificación se invierte para poder continuar con la validación.

Una vez que se determinó que la base de datos se encontraba libre de errores, se generó una nueva variable equivalente al resultado de la suma de cada uno de los reactivos de la escala, para posteriormente obtener los valores del percentil 25 y 75 de esta variable mediante un análisis de frecuencias. A partir de los valores obtenidos se creó una nueva variable dicotómica basada en los cuartiles extremos de la suma de los reactivos. Obtenida esta variable, se utilizó como variable de agrupación para realizar pruebas t de Student para evaluar la capacidad de discriminación de cada uno de los reactivos por comparación de grupos extremos, consiguiendo los resultados presentados en la tabla 2.

Tabla 2. Análisis de Discriminación de los reactivos de la escala de Persistencia y Perseverancia

	t	gl	Sig. (bilateral)
No abandono una tarea hasta que la he concluido	15.16	188.28	.000
Soy una persona que busca cumplir sus objetivos	22.08	130.74	.000
Sin importar los obstáculos, termino lo que inicio	20.88	136.52	.000
Me considero una persona muy trabajadora	17.18	147.85	.000
Cuando trabajo nunca me distraigo	5.87	198	.000
Nunca termino lo que comienzo	18.09	122.65	.000
Me rindo fácilmente	18.27	146.19	.000
Tiendo a cambiar mis objetivos y metas	11.92	169.46	.000
Dejo que las cosas salgan como sea	20.07	122.47	.000

Posteriormente para cada uno de los reactivos que discriminan, en este ejemplo fueron todos, se realizó el análisis de direccionalidad a través de tablas cruzadas con los grupos extremos. Con todos y cada uno de los reactivos que reúnen puntajes aprobatorios en los pasos recién mencionados, se realiza una prueba de confiabilidad interna mediante la fórmula de Alfa de Cronbach, donde junto con el estadístico de confiabilidad se revisan los valores de correlación de cada reactivo con el total, la correlación al cuadrado con los otros reactivos y el valor de modificación del estadístico de confiabilidad al eliminar el reactivo, obteniéndose un valor de Alfa de Cronbach (α) de .86. A continuación se presentan los estadísticos adicionales obtenidos.

Como se puede observar en esta Tabla con excepción del cuarto de los reactivos, Cuando trabajo nunca me distraigo, las correlaciones fueron altas; sin embargo, el α no se modifica sustancialmente para eliminarlo.

A continuación se realizó un análisis de correlación para los reactivos de la escala, usando la fórmula de Pearson para determinar el tipo de rotación a utilizar en el análisis factorial decidiendo debía ser ortogonal, ya que las correlaciones de Pearson resultantes fueron medianas (entre .3 y .7). Del gráfico de sedimentación (scree plot) y la matriz de componentes rotados resultante se obtuvo una estructura factorial compuesta por dos factores que en conjunto explican el 72.590% de la varianza. A continuación se presenta una tabla con la estructura factorial resultante así como los valores de la media, desviación estándar y Alfa de Cronbach por factor.

Tabla 3. Estadísticos de correlación y de eliminación de elemento del Alfa de Cronbach

	Correlación elemento-total corregida	Correlación múltiple al cuadrado	Alfa de Cronbach si se elimina el elemento
No abandono una tarea hasta que la he concluido	.636	.507	.846
Soy una persona que busca cumplir sus objetivos	.755	.684	.834
Sin importar los obstáculos, termino lo que inicio	.733	.688	.837
Cuando trabajo nunca me distraigo	.220	.183	.880
Nunca termino lo que comienzo	.583	.483	.851
Me rindo fácilmente	.609	.486	.849
Tiendo a cambiar mis objetivos y metas	.458	.396	.862
Dejo que las cosas salgan como sea	.703	.586	.839

Tabla 4. Estructura factorial para la escala de Persistencia y Perseverancia

	Factor	
	1	2
Media	3.54	3.48
Desviación estándar	1	1.14
Alfa de Cronbach	.85	.83
Me considero una persona muy trabajadora	.857	
Sin importar los obstáculos, termino lo que inicio	.815	
Soy una persona que busca cumplir sus objetivos	.770	
No abandono una tarea hasta que la he concluido	.724	
Cuando trabajo nunca me distraigo	.606	
Me rindo fácilmente		.797
Nunca termino lo que comienzo		.789
Dejo que las cosas salgan como sea		.780
Tiendo a cambiar mis objetivos y metas		.779

Método de extracción: Análisis de componentes principales con rotación ortogonal.
rotación ha convergido en 3 iteraciones.

El análisis de la composición factorial nos muestra claramente dos factores, el primero reflejando la persistencia y perseverancia en la tarea y el segundo negándola, con alta consistencia interna, aún más alta que la prueba original.

Discusión

El objetivo primordial de la presente aportación es promover el procedimiento metodológico y estadístico de validación psicométrica culturalmente relevante. A partir del ejemplo presentado anteriormente, se desprenden resultados que permiten asegurar que los índices de confiabilidad y validez pueden ser considerados como indicadores veraces de las propiedades psicométricas de la escala, lo que permitirá que esta escala pueda ser retomada para futuras aplicaciones con poblaciones similares.

Es conveniente resalta que, en este caso, puesto que las respuestas son objetivas, escala tipo Likert, no hubo necesidad de revisar los criterios de evaluación de las respuestas que también deben pasar por adaptación a la cultura tal y como propusimos hace más de cuatro décadas con estudios de adaptación de la Escala de Inteligencia para niños de Wechsler (Reyes Lagunes, 1965), En aquella ocasión, ejemplificando con el subtest de comprensión contrastándola con la automodificación, característica predominante en el mexicano, reportábamos que varios de los ejemplos que merecían, de acuerdo con la cultura norteamericana, un cero podrían ser consideradas como merecedoras de un mejor puntaje porque reflejan la solución del problema pero requieren de la autorización materna.

Obviamente, es indispensable después de la adaptación y validación obtener las normas específicas para la nueva población, recordemos que medimos a los individuos para compararlos al grupo al que pertenecen o al que quieren pertenecer.

Referencias

Díaz Guerrero, R. (1961). *Estudios de Psicología del Mexicano*. México: Ed. Antigua Librería Robledo.

International Personality Item Pool. (23 de Enero de 2008). *International Personality Item Pool.* Recuperado el 10 de Febrero de 2008, de http://ipip.ori.org/

Peterson, C., & Seligman, M. E. (2004). *Character Strengths and Virtues: A Handbook and Classification.* Oxford University Press.

Reyes Lagunes, I. (1965). *El Wechsler para niños en México. Consideraciones Psicológicas sobre adaptación.* Tesis de profesional no publicada. Universidad Nacional A[illegible]

SIGNIFICADO PSICOLÓGICO DEL PERDÓN Y LA RECONCILIACIÓN EN ESTUDIANTES UNIVERSITARIOS MEXICANOS

Irene Salas-Menotti[1]*, Alejandra Domínguez Espinosa**, Sonia Yuruen Lerma Mayer**

**Universidad Santo Tomás, Colombia y **Universidad Iberoamericana, México*

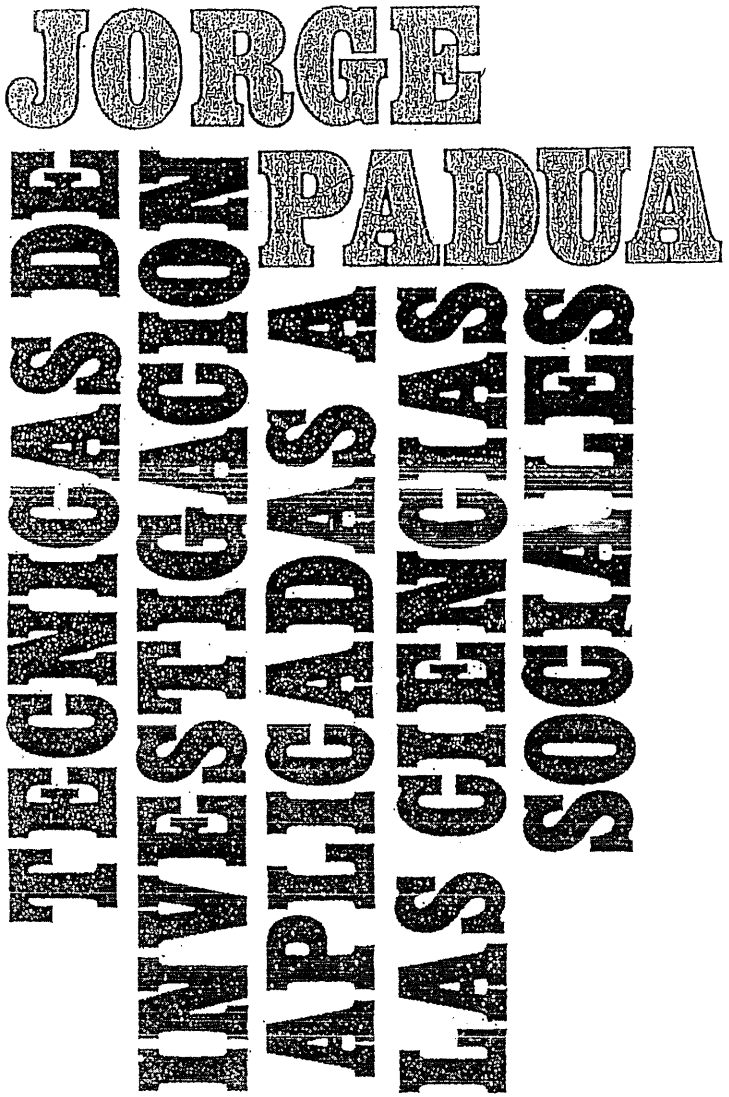
El perdón y la reconciliación son elementos importantes en la solución de conflictos, es por esto que difícilmente se puede dar por terminado un conflicto, sin el concurso de la capacidad de perdonar y la reconciliación. Perdonar a quién nos ha infligido una ofensa o perjuicio significa comprender lo que ha sucedido, el cómo y el porqué y por tanto representa la liberación de las limitaciones que implican el rencor, el odio, el temor y los deseos de venganza, que tienen un impacto determinante en el bienestar subjetivo de las personas. La relevancia del perdón, es que es utilizado constantemente en el lenguaje de la solución de conflictos de todo tipo: de pareja, familiares, entre amigos, entre pueblos, países, etc. Los jóvenes son el grupo de edad más susceptible a presentar conflictos interpersonales, por encontrarse en un momento del ciclo vital en donde están reafirmando su identidad y expectativas, lo cual puede generar conflictos que deben ser resueltos de manera eficaz. Es de vital importancia corroborar si las nociones teóricas del perdón son congruentes con la significación real que existe de ellos, si no es así, cualquier intento de intervención estará condenado al fracaso.

A pesar de la gran cantidad e información teórica y empírica de los conceptos de perdón y reconciliación, existe un gran vacío en cuanto a los significados o representaciones que los mexicanos tienen sobre éstos; se asume que todos los actores de un conflicto entienden lo mismo, y que es un concepto universal, como el de Justicia o Igualdad, cuando está plenamente demostrado que el contexto histórico-socio-cultural, así como aspectos circundantes como los valores, las creencias, las actitudes, los prejuicios y estereotipos, influyen directamente en los significados que se tienen sobre los fenómenos que revisten importancia para los miembros de una comunidad (Salas-Menotti y Reyes Lagunes, 2003).

Desde tiempos inmemoriales el perdón ha sido utilizado en el discurso que regula la interacción de los seres humanos, busca ordenar, reglamentar y restituir las relaciones entre personas que han sido agraviadas; sin embargo el perdón no sólo se puede entender desde sus aspectos jurídicos o formales, ya que está directamente asociado con la esfera de las creencias e ideologías religiosas de las personas, aspecto que de ninguna manera puede ser ignorado, ya que

[1] irenesalasmenotti@gmail.com

Padua, J. (1979). *Técnicas de investigación aplicadas a las ciencias sociales.* D.F.: El Colegio de México y El Fondo de Cultura Económica. pp.154-230



VI. ESCALAS PARA LA MEDICIÓN DE ACTITUDES

JORGE PADUA
INGVAR AHMAN

EL PROPÓSITO de este capítulo es presentar una serie de escalas —conocidas como escalas para la medición de actitudes— pero que de hecho pueden ser utilizadas para mediciones de otras variables.

La idea principal es la de posibilitar al lector la construcción de cada una de ellas, indicando los diferentes pasos con los mayores detalles. Como en los capítulos anteriores, recomendamos al lector interesado en el tema recurrir a la bibliografía que detallamos al final del capítulo.

Previamente a la presentación de cada tipo de escala conviene realizar una serie de aclaraciones que son importantes tanto para la construcción de las escalas, como para la interpretación de sus resultados:

1) La primera disquisición tiene que ver con el problema de la *medición*. Si bien en otro capítulo de este libro presentamos los diferentes niveles de medición en forma más extensa, interesa aquí dar un resumen sobre lo que se entiende por "medir" en ciencias sociales. La medición, de hecho, corresponde a una serie de teorías conocidas como *niveles de medición*: detrás de cada uno de los distintos niveles de medición están operando una serie de principios logicomatemáticos, que van a determinar o no el isomorfismo entre un concepto y el nivel de medición apropiado; es decir, el problema de la medición es expresado aquí como un problema en el cual se busca que el modelo matemático sea isomórfico con el concepto; esto es, que la "forma" del modelo sea idéntica a la "forma" del concepto. Si esto no ocurre (recuérdese que no hay isomorfismos "parciales"), estamos deformando el concepto (por ejemplo, si aplicamos niveles o modelos matemáticos intervalares a conceptos solamente operacionalizados a nivel ordinal es muy probable que los resultados ulteriores sean una consecuencia más del modelo matemático que del concepto en *sí*). Los niveles de medición más utilizados en ciencias sociales son:

a) Nivel nominal: la operación de medir consiste simplemente en la asignación de nombres o de números a distintas categorías. La función del número en este nivel de medición es muy elemental, ya que simplemente sirve para distinguir diferentes categorías. A este nivel se está "midiendo" cuando por ejemplo se hace la distinción entre varón y mujer o entre católico, protestante, judío, mahometano, otra religión. La operación de medición consistiría entonces en referir la observación a una clase o categoría, para luego contar cuántas frecuencias caen dentro de cada categoría. Uno no puede hacer legítimamente ninguna afirmación que vaya más allá de las diferentes distinciones.

b) Nivel ordinal: en este nivel de medición uno está en condiciones de distinguir entre diferentes categorías y de poder afirmar si una categoría posee

en *mayor*, *menor* o *igual* grado el atributo que estamos midiendo. La escala de jerarquía militar en el ejército es un buen ejemplo: un sargento tiene *menos* autoridad y es *diferente* que un teniente; éste a su vez tiene *menor* grado de autoridad que un capitán. Uno puede ordenar entonces las categorías sargento, teniente, capitán, con respecto a autoridad de la siguiente manera:

Capitán > Teniente > Sargento

Incluso puede llegarse a establecer comparaciones con distribuciones de autoridad para distintas categorías en la marina o las fuerzas aéreas, etcétera.

c) Nivel intervalar: existe aún mayor precisión que en los anteriores, ya que no solamente podemos categorizar y establecer relaciones de mayor, menor o igual, sino además calcular la distancia entre los intervalos o categorías. Obsérvese que en el caso de las mediciones de nivel ordinal uno puede afirmar que un sargento tiene menos autoridad que un teniente, y que un capitán tiene más autoridad que éste; sin embargo, no estamos en condiciones de precisar *cuánto más* o *cuánto menos*. Mediante la adjudicación de un cero arbitrario en esta escala podemos especificar la distancia entre esas categorías.

d) Nivel racional: es la forma de medición que utiliza valores cero absolutos, y que nos permiten establecer diferencias entre cualquier par de objetos a un máximo de precisión. A esta escala pertenecen el sistema métrico y el de pesos por ejemplo. La diferencia entre el nivel intervalar y el nivel racional, es que por ejemplo en la medición de la temperatura en que se utiliza escala intervalar, no es posible afirmar correctamente que cuando se registran 40 grados de temperatura estamos sintiendo el *doble* de calor que cuando teníamos 20 grados; por el contrario, si nos desplazamos 40 kilómetros en línea recta sí podemos afirmar que hemos duplicado la distancia recorrida al desplazarnos 20 kilómetros.

La mayor parte de las escalas de medición de actitudes que vamos a describir se encuentran comprendidas entre la escala ordinal y la escala intervalar.

2) La segunda disquisición a considerar también ha sido tratada en otro capítulo. Tiene que ver con la dimensionalización de los conceptos. (Ver "El proceso de investigación".) Recordemos solamente que muchas de las variables con las que trabajan los científicos sociales son complejas y están compuestas de una serie de dimensiones o atributos. Por ejemplo, la variable "religiosidad" puede ser concebida como compuesta de tres dimensiones: dogmatismo, misticismo, ritualismo, cada una de las cuales tiene distintos indicadores. O la del *status* socioeconómico, que tradicionalmente se dimensionaliza en ocupación, educación e ingreso. El investigador en general espera que las dimensiones estén intercorrelacionadas, es decir, uno espera que una persona con alto grado de religiosidad manifieste valores altos en las dimensiones ritualismo, misticismo y dogmatismo. Idénticamente en el caso del *status* socioeconómico es de esperar que la educación se correlacione fuertemente con la ocupación y el ingreso. Sin embargo, existen casos en los cuales se producen "inconsistencias": sujetos con alta educación y bajo ingreso; alta ocupación y

VI. ESCALAS PARA LA MEDICIÓN DE ACTITUDES

JORGE PADUA
INGVAR AHMAN

EL PROPÓSITO de este capítulo es presentar una serie de escalas —conocidas como escalas para la medición de actitudes— pero que de hecho pueden ser utilizadas para mediciones de otras variables.

La idea principal es la de posibilitar al lector la construcción de cada una de ellas, indicando los diferentes pasos con los mayores detalles. Como en los capítulos anteriores, recomendamos al lector interesado en el tema recurrir a la bibliografía que detallamos al final del capítulo.

Previamente a la presentación de cada tipo de escala conviene realizar una serie de aclaraciones que son importantes tanto para la construcción de las escalas, como para la interpretación de sus resultados:

1) La primera disquisición tiene que ver con el problema de la *medición*. Si bien en otro capítulo de este libro presentamos los diferentes niveles de medición en forma más extensa, interesa aquí dar un resumen sobre lo que se entiende por "medir" en ciencias sociales. La medición, de hecho, corresponde a una serie de teorías conocidas como *niveles de medición*: detrás de cada uno de los distintos niveles de medición están operando una serie de principios logicomatemáticos, que van a determinar o no el isomorfismo entre un concepto y el nivel de medición apropiado; es decir, el problema de la medición es expresado aquí como un problema en el cual se busca que el modelo matemático sea isomórfico con el concepto; esto es, que la "forma" del modelo sea idéntica a la "forma" del concepto. Si esto no ocurre (recuérdese que no hay isomorfismos "parciales"), estamos deformando el concepto (por ejemplo, si aplicamos niveles o modelos matemáticos intervalares a conceptos solamente operacionalizados a nivel ordinal es muy probable que los resultados ulteriores sean una consecuencia más del modelo matemático que del concepto en sí). Los niveles de medición más utilizados en ciencias sociales son:

a) Nivel nominal: la operación de medir consiste simplemente en la asignación de nombres o de números a distintas categorías. La función del número en este nivel de medición es muy elemental, ya que simplemente sirve para distinguir diferentes categorías. A este nivel se está "midiendo" cuando por ejemplo se hace la distinción entre varón y mujer o entre católico, protestante, judío, mahometano, otra religión. La operación de medición consistiría entonces en referir la observación a una clase o categoría, para luego contar cuántas frecuencias caen dentro de cada categoría. Uno no puede hacer legítimamente ninguna afirmación que vaya más allá de las diferentes distinciones.

b) Nivel ordinal: en este nivel de medición uno está en condiciones de distinguir entre diferentes categorías y de poder afirmar si una categoría posee en *mayor, menor* o *igual* grado el atributo que estamos midiendo. La escala de jerarquía militar en el ejército es un buen ejemplo: un sargento tiene *menos* autoridad y es *diferente* que un teniente; éste a su vez tiene *menor* grado de autoridad que un capitán. Uno puede ordenar entonces las categorías sargento, teniente, capitán, con respecto a autoridad de la siguiente manera:

Capitán > Teniente > Sargento

Incluso puede llegarse a establecer comparaciones con distribuciones de autoridad para distintas categorías en la marina o las fuerzas aéreas, etcétera.

c) Nivel intervalar: existe aún mayor precisión que en los anteriores, ya que no solamente podemos categorizar y establecer relaciones de mayor, menor o igual, sino además calcular la distancia entre los intervalos o categorías. Obsérvese que en el caso de las mediciones de nivel ordinal uno puede afirmar que un sargento tiene menos autoridad que un teniente, y que un capitán tiene más autoridad que éste; sin embargo, no estamos en condiciones de precisar *cuánto más* o *cuánto menos*. Mediante la adjudicación de un cero arbitrario en esta escala podemos especificar la distancia entre esas categorías.

d) Nivel racional: es la forma de medición que utiliza valores cero absolutos, y que nos permiten establecer diferencias entre cualquier par de objetos a un máximo de precisión. A esta escala pertenecen el sistema métrico y el de pesos por ejemplo. La diferencia entre el nivel intervalar y el nivel racional, es que por ejemplo en la medición de la temperatura en que se utiliza escala intervalar, no es posible afirmar correctamente que cuando se registran 40 grados de temperatura estamos sintiendo el *doble* de calor que cuando teníamos 20 grados; por el contrario, si nos desplazamos 40 kilómetros en línea recta sí podemos afirmar que hemos duplicado la distancia recorrida al desplazarnos 20 kilómetros.

La mayor parte de las escalas de medición de actitudes que vamos a describir se encuentran comprendidas entre la escala ordinal y la escala intervalar.

2) La segunda disquisición a considerar también ha sido tratada en otro capítulo. Tiene que ver con la dimensionalización de los conceptos. (Ver "El proceso de investigación".) Recordemos solamente que muchas de las variables con las que trabajan los científicos sociales son complejas y están compuestas de una serie de dimensiones o atributos. Por ejemplo, la variable "religiosidad" puede ser concebida como compuesta de tres dimensiones: dogmatismo, misticismo, ritualismo, cada una de las cuales tiene distintos indicadores. O la del *status* socioeconómico, que tradicionalmente se dimensionaliza en ocupación, educación e ingreso. El investigador en general espera que las dimensiones estén intercorrelacionadas, es decir, uno espera que una persona con alto grado de religiosidad manifieste valores altos en las dimensiones ritualismo, misticismo y dogmatismo. Idénticamente en el caso del *status* socioeconómico es de esperar que la educación se correlacione fuertemente con la ocupación y el ingreso. Sin embargo, existen casos en los cuales se producen "inconsistencias": sujetos con alta educación y bajo ingreso; alta ocupación y

baja educación, etc.; o en el caso de religiosidad sujetos que responden positivamente a *items* de dogmatismo y negativamente a *items* de ritualismo. En el caso de escalas es posible construir *escalas multidimensionales o escalas unidimensionales.*

3) La tercera digresión es parte ya del discurso propio de la construcción de escalas para la medición de actitudes y tiene que ver con los modos en que se incluyen o eliminan los *items* de una escala. Una vez que los *items* o los juicios de actitud han sido formulados o recolectados, los métodos utilizados para su inclusión en las escalas son:

a) Uso de jueces que no responden a los *items* en términos del grado de acuerdo o desacuerdo que se tenga con ellos, sino en el grado de validez que el juez otorgue al *item* o juicio en relación a la variable. Es decir que los jueces son utilizados aquí para determinar el valor que el investigador va a asignar al *item* sobre un continuo psicológico. Una vez que los juicios o *items* tienen asignado un valor, se aplican a los sujetos para que ellos *sí* expresen su grado de acuerdo o desacuerdo. Los puntajes definitivos serán computados a partir del valor dado por los jueces. Las escalas que desarrollaremos más adelante y que utilizan este sistema son: la del método de intervalos aparentemente iguales de Thurstone, la de intervalos sucesivos y la de comparación por pares.

b) El método de las respuestas directas con los *items* o juicios. Este método no requiere el conocimiento previo de los valores de escala, sino que los puntajes se determinan en función de las respuestas. No es necesaria la utilización de jueces en el sentido expresado en el parágrafo anterior. El método a examinar aquí será el del análisis de escalograma y el de la escala Lickert.

c) Finalmente, en la combinación de método de respuesta y de uso de jueces, el método a examinar será el de la técnica de la escala discriminatoria (*scale-discrimination technique*).

4) La cuarta digresión tiene que ver con los problemas de confiabilidad y validez, y a que los resultados en el *test* pueden sufrir variaciones de tres tipos:

a) Variaciones en el instrumento: los errores producto de instrumentos "mal calibrados" (problema de validez).

b) Variaciones en los sujetos: dicho en otros términos, en dos aplicaciones distintas en el tiempo, el sujeto produce resultados distintos (problemas de confiabilidad).

c) Variaciones simultáneas en los sujetos y en el instrumento: por instrumento "válido" entendemos aquel que mide efectivamente lo que se propone medir; mientras que por "confiable" entendemos que mide siempre de la misma manera.

5) La quinta digresión se relaciona estrechamente con la tercera y la cuarta, y se refiere a las formas en que sean clasificadas las escalas según estén centradas en los sujetos, en el instrumento o en ambos:

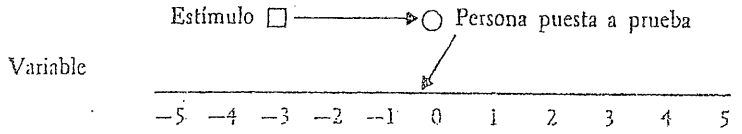
a) El enfoque centrado en el instrumento, llamado *stimulus-centered approach* (Torgerson), es aquel en el cual la variación sistemática en la reacción de los sujetos al estímulo es atribuida a diferencias en éste.

b) En el enfoque centrado en el sujeto (*subject-centered approach*) la variación al estímulo es atribuida a diferencias individuales en los sujetos.

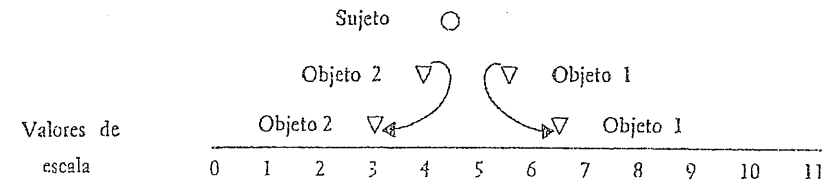
c) En el enfoque centrado en la respuesta (*response approach*) la variabilidad de reacciones al estímulo se atribuye tanto a las reacciones de los sujetos como al instrumento.

6) Otra forma de clasificar a los instrumentos o escalas, es la propuesta por Coombs,[1] de *tests* de Tipo A y *tests* de Tipo B:

a) El *test* de tipo A: sirve para determinar algunas propiedades de un sujeto o de un objeto en el medio de una persona. A través de su conducta en la situación de prueba (es decir, su rendimiento en un *item* específico de la escala) el sujeto, consciente o inconscientemente se sitúa en una posición a lo largo del continuo en la variable que la escala está midiendo. Gráficamente:



b) El *test* de tipo B: se usa para determinar las propiedades de un objeto. El sujeto indicaría entonces la posición de uno o más objetos (*items* en nuestro caso) a lo largo de la variable. En la construcción de escalas, el *test* de tipo B se utiliza por ejemplo cuando se pide a jueces que ubiquen el valor que un determinado *item* tiene en cierta variable (ver escala Thurstone). La situación gráfica es la siguiente:



Ejemplo: Si usted le pregunta a un número de personas cuál es el partido político que tiene el programa más conservador en Argentina, el *item* o pregunta puede ubicarse como un *test* de tipo B. Si la pregunta fuera: ¿Cuál es el partido con el que usted más simpatiza?, se trataría de un *test* de tipo A.

7) La última digresión tiene que ver con *los items*. Por ahora podemos definir al *item* como una frase o juicio indicador de la variable que estamos tra-

[1] Coombs, C. H.: *A Theory of Psychological Scaling*; Engineering Research Institute, University of Michigan; Ann Arbor, Mich., 1952.

tando de medir. La formulación del *item* en el caso que estemos trabajando con cuestionarios es siempre un juicio, nunca una pregunta o una interrogación. Para relacionar los *items* con los niveles de medición mencionados en la primera digresión tomemos los siguientes ejemplos:

A) ¿Entre cuáles de los siguientes grupos de altura se ubica usted? (marque con una cruz la que corresponda)

☐ 120-139 cm

☐ 140-159 cm

☐ 160-179 cm

☐ 180-199 cm

☐ 200-219 cm

B) Si usted tuviera que definir sus intereses en los asuntos políticos, ¿diría que está muy interesado, algo interesado, ni desinteresado ni interesado, desinteresado o muy desinteresado?

☐ Muy interesado

☐ Interesado

☐ Indiferente

☐ Desinteresado

☐ Muy desinteresado

C) De los siguientes, ¿cuál es el diario que lee con mayor frecuencia?

☐ *Excelsior.*

☐ *Novedades.*

☐ *Últimas Noticias.*

☐ *El Heraldo.*

☐ *El Día.*

☐ *El Sol de México.*

☐ *Ovaciones.*

☐ *La Prensa.*

☐ Otro: (especificar) .....

El ejemplo A señala una variable medida a nivel racional; el ejemplo B corresponde a un nivel de medición ordinal y el ejemplo C corresponde a un nivel de medición nominal.

Ahora bien, si tratamos de medir la altura de una persona, como en el caso del ejemplo A, es evidente que podemos hacer la medición de varias maneras. Por un lado podríamos tener diferentes varas con diferentes alturas: una vara de 175, otra de 176, otra de 177, etc. Supongamos que nos llegue un sujeto de 178 cms de altura; le aplicamos la vara de 160; la rechazamos diciendo que el sujeto es más alto, y seguimos aplicando varas hasta que llegamos a la correcta. En este caso estamos tratando con lo que se denomina *items acumulativos,* que en el cuestionario autoadministrado aparecerían de la siguiente forma: (las marcas aparecen para el caso de nuestro sujeto que mide 178 cm)

Item 1: ¿Tiene Ud. más de 170 cm de altura?	Si ☒	No ☐
Item 2: ¿Tiene Ud. más de 171 cm de altura?	Si ☒	No ☐
Item 3: ¿Tiene Ud. más de 172 cm de altura?	Si ☒	No ☐
Item 4: ¿Tiene Ud. más de 173 cm de altura?	Si ☒	No ☐
Item 5: ¿Tiene Ud. más de 174 cm de altura?	Si ☒	No ☐
Item 6: ¿Tiene Ud. más de 175 cm de altura?	Si ☒	No ☐
Item 7: ¿Tiene Ud. más de 176 cm de altura?	Si ☒	No ☐
Item 8: ¿Tiene Ud. más de 177 cm de altura?	Si ☒	No ☐
Item 9: ¿Tiene Ud. más de 178 cm de altura?	Si ☐	No ☒
Item 10: ¿Tiene Ud. más de 179 cm de altura?	Si ☐	No ☒
Item 11: ¿Tiene Ud. más de 180 cm de altura?	Si ☐	No ☒
Etcétera.		Etc.

Si nosotros conocemos que una persona ha contestado a una de las preguntas con un "Sí", sabemos que todas las preguntas que están por "debajo" tienen también, por derivación, una respuesta "Sí". De la misma manera, si una persona ha respondido "No", todas las preguntas que estén por "arriba" tendrán una respuesta similar.

Una alternativa diferente de presentar la pregunta es a través de *items diferenciados,*[2] en la que ante una serie de *items,* la persona entrevistada marcará solo una o algunas alternativas. En un cuestionario, la pregunta sería realizada de la siguiente manera:

[2] Un ejemplo mucho más claro sería preguntar en qué país nació, o cuál es su sexo.

Item 1: ¿Está Ud. entre los 170-174 cm de altura?	Si	(No)
Item 2: ¿Está Ud. entre los 175-179 cm de altura?	(Si)	No
Item 3: ¿Está Ud. entre los 180-184 cm de altura?	Si	(No)
Item 4: ¿Está Ud. entre los 185-190 cm de altura?	Si	(No)

Un *item* incluido en una serie de *items* diferenciados debe ser formulado de manera tal que sea respondido con un "Sí" únicamente por las personas que tienen una posición fija a lo largo de la variable. Un *item* incluido en una serie de *items* acumulativos, debe ser formulado de manera tal que solamente pueda ser respondido con un "Sí" a un lado de una determinada posición a lo largo de la variable investigada.

Trabajando con la altura de las personas, la diferencia entre estos dos tipos de serie de *items* aparece como trivial. Sin embargo, en el tratamiento de las escalas para medición de actitudes, su diferenciación es importante. Consecuentemente, en cada tipo de las escalas a examinar (Lickert, Thurstone, Guttman, comparación por pares) tomaremos en cuenta la especificación de esta perspectiva.

Escalas para la medición de actitudes

Vamos a utilizar la definición de escalas hecha por Stoufer:[3]

> Se dice que existe escala cuando a partir de una distribución de frecuencias multivariada de un universo de atributos, es posible derivar una variable cuantitativa con la cual caracterizar los objetos de un modo tal que cada atributo sea una función simple de aquella variable cuantitativa.

Una escala es una forma particular de *índice,* aunque aquí utilizaremos una serie de procedimientos objetivos para la selección de *items* de manera tal de controlar los errores producto de la subjetividad del investigador. (Ver el capítulo sobre "Conceptos, indicaciones e índices".) Construir una escala implica una serie de procedimientos mediante los cuales —de acuerdo a distintas reglas— se seleccionan *items* y se adjudican números a un conjunto de *items* (juicios o sentencias), número que va a expresar la intensidad que un sujeto o grupo de sujetos manifiestan en la variable.

Las actitudes en el contexto individual representan un estado mental que es un puente entre estados psicológicos y objetos exteriores. Kretch y Cruschfield [4] sostienen a este respecto que se puede definir a una actitud "como una organización durable de procesos motivacionales, emocionales, perceptuales y cognitivos con respecto a algún aspecto del mundo del individuo". Las actitudes serían entonces procesos claves para entender las tendencias del individuo en relación con objetos y valores del mundo externo, aun que esas tendencias no son estáticas y, como las de Newcomb,[5] las actitudes representan un residuo de la experiencia anterior del sujeto. Las actitudes perdurarían en el sentido que tales residuos "son trasladados a nuevas situaciones, pero cambian en la medida en que nuevos residuos son adquiridos a través de experiencias en situaciones nuevas".

Las actitudes serían entonces tendencias a actuar con respecto a alguna entidad especificable (Newcomb); o como lo quieren Thomas y Znaniecki: [6] "la tendencia individual a reaccionar, positiva o negativamente, a un valor social dado".

Las actitudes medidas por escalas deben interpretarse en términos analíticos no como "hechos", sino como "síntomas". Existe una serie de conceptos relacionados a las actitudes; entre ellos detallamos los siguientes:

Creencia: actitudes que incorporan una cantidad importante de estructuración cognitiva. Las actitudes son *hacia* algo, mientras que las creencias son *en* o *sobre* algo.

Sesgo (bias): son actitudes o prejuicios débiles, basados en premisas incompletas, deducidas falsamente o preconcebidas. Por lo tanto son poco precisas y relativamente fáciles de cambiar.

Doctrina: son objetos estímulos elaborados, hacia los cuales el individuo manifiesta una actitud. Una doctrina (republicana, liberal, católica, comunista, etc.) describe específicamente las razones para adherencia; por lo tanto una doctrina se aprende.

Fe: implica una actitud con alta carga emocional o afectiva. El sistema de actitudes referentes a la fe, describe una creencia fundamental y específica de la persona. La fe se ubica entre la creencia y la ideología.

Ideología: es un sistema cognitivo elaborado, que sirve para justificación de formas específicas de comportamiento, o como medio de racionalización. Ideología es concebido como un sistema lógico falso. La ideología se acepta como una fe.

Valor: en un sentido psicológico amplio, valores son marcos de referencia que sirven de guía o mapa para la evaluación de la experiencia y la conducta. Sistema de valores, sería la organización elaborada y articulada de actitudes, que toman valencias positivas, negativas o neutras en cuanto a objetos, estímulos, del ambiente y la relación de éstos con las metas de vida.

Opinión: son evaluaciones tentativas, no fijas, sujetas a cambio o inversión. Es decir, son menos fijas y no comprometedoras para el individuo.

[3] Stoufer, S. *et al.*: *Studies in Social Psychology in World War II. Measurement and prediction* (vol. IV); John Wiley & Sons, Nueva York, 1966.

[4] Kretch D.; Cruschfield, R. S.: *Theory and Problems of Social Psychology;* McGraw-Hill, Nueva York, 1948.

[5] Newcomb, T. M.: *Personality and Social Change: attitude formation in a Student community;* Dryden, Nueva York, 1943.

[6] Thomas, W. I. y Znaniecki, F.: *The Polish peasant in Europe and America:*

ACTITUDES E INTERESES: ALGO ACERCA DE SU DIRECCIÓN E INTENSIDAD

Señalamos brevemente aquí algunas propiedades de las actitudes que tienen que ver con los problemas asociados a la medición de las mismas a través de escalas.

1) Dirección. La actitud puede ser positiva o negativa. En algunos casos es explicable hablar de una actitud neutral cuando el sujeto no es ni positivo ni negativo hacia el objeto.

2) Intensidad. La intensidad de la actitud es alta si el sujeto está fuertemente convencido que la actitud es justificada y baja si el sujeto no piensa así. Al respecto, una medida aceptable es la frecuencia con la cual el sujeto usa fuertes expresiones que señalan un engrane emocional y en la medida en que él está preparado a argumentar en favor de su posición.

3) Estabilidad. La actitud es estable si permanece invariable por un periodo muy largo.

4) Fortaleza. Una actitud es fuerte si, por ejemplo, es difícil de cambiar a través de persuasión o propaganda, y suave si cambia fácilmente.

5) Importancia. La importancia de la actitud es grande, si influye sobre la conducta de una persona en muchos campos de actividades.

6) Visibilidad, observabilidad o relevancia externa. Ésta es alta si es fácil concluir a partir de observaciones sobre la conducta del sujeto (manera de hablar, acciones) que él tiene cierta actitud.

7) Relevancia interna. Es detectable si en la experiencia del sujeto la actitud por sí misma tiene una parte importante en su mundo de referencia. La actitud de una persona hacia la religión puede tener baja relevancia externa y, al mismo tiempo, una alta relevancia interna.

8) Involucramiento del ego. Cuando una actitud llega a constituir una parte importante de la personalidad, o es algo adherido a ella.

9) Integración y aislamiento. De un sistema mayor de actitudes interrelacionadas (perspectiva de vida, ideologías). Las actitudes políticas o religiosas están frecuentemente incluidas en tal sistema complejo, mientras que una actitud hacia una cierta pasta de dientes es experimentada como aislada, sin relación lógica o psicológica al complejo sistema total.

10) Especificidad o precisión. La imaginación es dirigida hacia la actitud. Ejemplo: un profesor de teología tiene una actitud más específica hacia la Iglesia católica que el "hombre de la calle".

11) Verificabilidad. La imaginación es igualmente dirigida hacia la actitud. El conocimiento es verificable, la fe, no; las opiniones ocupan un lugar intermedio y pueden a veces ser verificadas.

Debe notarse que muchas de las propiedades de las actitudes están correlacionadas. Alta intensidad frecuentemente significa estabilidad, fortaleza, alta relevancia interna y externa y alto grado de involucramiento del ego. Es importante observar los casos en los cuales las propiedades no están combinadas. En dichos casos, las distinciones hechas anteriormente son obvias.

Las pruebas ordinarias de actitud sólo admiten la medición de algunas

de las propiedades mencionadas, principalmente dirección e intensidad. Para determinar las otras propiedades de la actitud podría ser útil una entrevista profunda del tipo entrevista focalizada.

Dirección e intensidad frecuentemente son vistas como una sola propiedad de la actitud. Este enfoque es justificado por lo siguiente: suponiendo que dirección e intensidad son dos variables independientes, si de cada una de ellas pueden obtenerse tres valores diferentes, las combinaciones siguientes son posibles:

CUADRO 1

Intensidad \ *Dirección*	*Positivo*	*Neutral*	*Negativo*
Fuerte	+++	000	− − −
Ni fuerte ni débil	+	00	− −
Débil		0	

Una serie de investigaciones (entre otras la realizada por Edward H. Suchman) han mostrado, sin embargo, que una actitud neutral demuestra una baja intensidad. En la mayoría de los casos, las dos áreas rayadas podrían ser consideradas como inexistentes.

Los resultados son siete combinaciones, las cuales pueden ser ordenadas de la siguiente manera:

— — — O — — —

Fuerte positivo — Neutral — Fuerte negativo

Esta perspectiva ha sido también usada en la construcción de un significativo "punto cero" para las escalas de actitudes. El Cuadro 2 resume la semejanza y diferencias entre tres de las principales escalas y el tipo de enfoque utilizado en ellas. En la parte final del capítulo agregamos el método de comparación por pares, de diferencial semántico y de escala de distancia social de Bogardus.

LA ESCALA LICKERT

Es un tipo de escala aditiva que corresponde a un nivel de medición ordinal, consistente en una serie de *ítems* o juicios ante los cuales se solicita la reacción del sujeto. El estímulo (*ítem* o sentencia) que se presenta al sujeto representa la propiedad que el investigador está interesado en medir y las respuestas son solicitadas en términos de grados de acuerdo o desacuerdo que el sujeto tenga con la sentencia en particular.

CUADRO 2. *Semejanzas y diferencias entre algunos tipos de escalas para la medición de actitudes y en los enfoq*
Centrado en el sujeto, Centrado en el estímulo y Response-approach

Semejanzas	*DIFERENCIAS*		
	Centrado en el sujeto	*Centrado en el estímulo*	*Response-approach*
Los tres tipos de enfoques se usan en la construcción de escalas que permiten medir distinciones de grado, más que de cualidad entre individuos.	Consiste en preguntarle a un sujeto su opinión acerca de un objeto para que él se ubique en determinado punto de la escala.	Consiste en preguntarle dónde ubicaría un estímulo sobre un continuo que representa el objeto de estudio.	El propósito puede ser escalar los sujetos, los estímulos, o ambos en relación a un objeto.
	Se dan valores a los sujetos.	Se dan valores a los estímulos.	Se dan valores tanto a los sujetos como a los estímulos.
	Se subrayan las diferencias individuales.	Se tratan de eliminar las diferencias individuales.	Se tienen en cuenta las diferencias en los sujetos y en los individuos.
	Análisis de la varianza unidireccional.	Análisis de la varianza unidireccional.	Análisis de la varianza en dos direcciones.
Semejanzas en las escalas Lickert-Thurstone-Guttman	*Ejemplo: escala Lickert*	*Ejemplo: escala Thurstone*	*Ejemplo: escala Guttman*
Con pequeñas modificaciones los mismos *items* pueden utilizarlos en los diferentes tipos de escala, ya que la	Escala ordinal.	Pretende ser intervalar.	Escala ordinaria.
	Escala aditiva.	Escala diferencial.	Escala acumulativa.
diferencia entre las escalas no reside en la elección de los *items*, sino en la relación lógica existente entre ellos.	Pretende ser unidimensional, pero de hecho se mezclan dimensiones.	Idem. a Lickert.	Es unidimensional, ésta es determinada por el coeficiente de reproductividad.
Las tres emplean el método de jueces en la construcción de la escala.	Los *items* se elaboran unidireccionalmente, procurando un 50 % de *items* positivos y un 50 % de *items* negativos.	Los *items* se formulan de manera que a sean positivos; a negativos y a neutros.	Como es una técnica de análisis, los *items* se formulan de acuerdo a la técnica que se decida utilizar, Lickert o Thurstone.
Estrictamente todas las escalas miden a nivel ordinal.	Los *items* son acumulativos, por ello pueden estar aproximadamente en la misma posición de la escala.	Los *items* son diferenciadores.	Cada *item* tiene el carácter de acumulativo.
En principio las tres son unidimensionales.			
En el momento de aplicar la escala ya validada (versión final), aunque se hayan utilizado distintos enfoques en la construcción, todos los procedimientos van a estar centrados en el sujeto. Lo que subyace en todos los procedimientos es el obtener conocimientos de las actitudes que forman parte de un contexto sociocultural.	La cantidad de *items* en la versión de los jueces varía de 30 a 50. En la versión final quedan entre 15 y 25.	El número de *items* iniciales puede llegar a los 200. En la escala final pueden quedar un máximo de 50.	La cantidad de *item* depende de la técnica de construcción.
	La selección de los *items* se realiza en base a su poder discriminativo.	La selección de los *items* se hace en base al recorrido intercuartílico.	La selección de los *items* se basa en su escalabilidad, que se determina en base al número de errores.
	Las alternativas de respuesta en cada *item* puede variar de 3, 4, 5 o más alternativas.	El puntaje de los *items* se determina por el valor de la mediana de los puntajes asignados por jueces, sobre una escala que va de 1 a 11.	El puntaje de los *items* varía de acuerdo a la técnica empleada.
	No existe gradación de los *items* a lo largo de un continuo.	Se presentan mezclados *items* positivos, negativos y neutros.	Los *items* se ordenan en forma decreciente de acuerdo a su grado de dificultad.

DIFERENCIAS		
Centrado en el sujeto	*Centrado en el estímulo*	*Response-approach*
El número de jueces varía de 50 a 100. Éstos deben tener características similares a las de los sujetos en el universo a estudiar.	El número de jueces es entre 50 y 200. Se exige de ellos objetividad e información.	El número de jueces depende de la técnica empleada en la construcción.
La consistencia interna de la escala se establece mediante el método de correlación por mitades (*split-half*).	La consistencia interna se basa en el cálculo de la correlación entre cada *ítem* con el puntaje total del *test* (*item-test*).	La consistencia interna está determinada por la escalabilidad, calculada en base a la diferencia CR-MMR.
A cantidad igual de *items* es más confiable que la escala Thurstone. La confiabilidad aumenta con el incremento en las alternativas de respuesta.	Cuando se alcanza un número de 50 *items* es más confiable que la Lickert.	Es más confiable debido a su unidimensionalidad.
Es más fácil y rápida de construir.	Difícil de construir. Gasto fuerte en términos de tiempo y trabajo.	Depende de la técnica utilizada en la construcción.
Escala correspondiente a *test* de Tipo A.	Corresponde a un *test* de Tipo B.	Corresponde a ambos tipos de *test* (A y B).
En la escala final se presentan los *items* con la cantidad de alternativa idéntica a la de la versión de los jueces.	En la escala final se presentan solamente dos alternativas de respuesta: acuerdo-desacuerdo.	En la escala final se presentan los *items* en orden de dificultad creciente.
El puntaje máximo es igual al número de *items* multiplicado por el puntaje mayor en cada alternativa de respuesta: el puntaje mínimo es igual al número de *items* multiplicado por el puntaje menor en las alternativas de respuesta. Para la ubicación de individuos se pueden utilizar también en valores promedio.	El puntaje máximo y mínimo en la escala se determinan por la sumatoria de *items* ponderados.	Los sujetos se ubican en la escala en forma decreciente: desde los que contestaron en forma positiva a todas las preguntas, hasta los que contestaron en forma negativa a todas las preguntas.
Con frecuencia la puntuación total de un individuo puede tener un significado poco claro, cuando se lo compara con otros individuos, ya que combinaciones distintas pueden producir el mismo resultado.	Idem. a la Lickert.	Los individuos que tienen un puntaje de actitud más favorable, deben también tener una actitud más favorable en cada *ítem* escalado.
El problema principal de la escala es la validez. Determinar cuándo una misma puntuación alcanzada por combinación de distintas categorías de respuesta tiene consecuencias para la interpretación del atributo en cuestión y cuándo no.	El problema principal es el centro de la escala. Éste no nos dice mucho acerca del significado que tiene el hecho que un individuo ocupe una posición en el centro de la escala.	En el caso de actitudes complejas, no es muy eficaz.

El método del *summated ratings* de Lickert resulta de la suma algebraica de las respuestas de individuos a *items* seleccionados previamente como válidos y confiables. Si bien la escala es aditiva, no se trata de encontrar *items* que se distribuyan uniformemente sobre un continuo "favorable-desfavorable", sino que el método de selección y construcción de la escala apunta a la utilización de *items* que son definitivamente favorables o desfavorables con relación al objeto de estudio. El puntaje final del sujeto es interpretado como su posición en una escala de actitudes que expresa un continuo con respecto al objeto de estudio.

La escala Lickert es pues un *test de Tipo A*, ya que el sujeto, a través de su conducta en la situación de prueba, consciente o inconscientemente, se sitúa a lo largo de la variable. La escala Lickert es también una escala del tipo *centrada en el sujeto* (*subject-centered*): el supuesto subyacente es que la variación en las respuestas será debida a diferencias individuales en los sujetos. Veremos además cuando examinemos los pasos en la construcción de la escala, que la escala inicial se administra a una muestra de sujetos que actuarán como *jueces*. (Esa muestra de sujetos debe ser representativa de la población a la que se aplicará la escala final.)

Finalmente, los *items* son seleccionados en base a su *poder discriminatorio* entre grupos con valores altos y con valores bajos en la variable. Es decir, que lo que interesa es la coherencia, entendida ésta en función de las respuestas

A) *La construcción de una escala Lickert*

La construcción de una escala de este tipo implica los siguientes pasos: *1)* necesario construir una serie de *items* relevantes a la actitud que se qui medir. *2)* Los *items* deben ser administrados a una muestra de sujetos c van a actuar como *jueces*. *3)* Se asignan puntajes a los *ítems* según la dir ción positiva o negativa del *item*. *4)* Se asignan los puntajes totales a sujetos de acuerdo al tipo de respuesta en cada *item*, la suma es algebrai *5)* Se efectúa un análisis de *items*. *6)* Se construye con base en los *items* leccionados la escala final. Examinemos en detalle cada uno de los pasos:

1) *La construcción de los* items

Los *items* que van a aplicarse a la muestra inicial de jueces, cuyo núme debe ser entre 30 y 50. Para la construcción de los *items* deben tomarse cuenta los siguientes criterios, que aparecen en Edwards:

a) Evite los *items* que apuntan al pasado en lugar del presente.

b) Evite los *items* que dan demasiada información sobre hechos, o aquellos que pueden ser interpretados como tales.

c) Evite los *items* ambiguos.

d) Evite los *items* irrelevantes con respecto a la actitud que quiere medir.

e) Los *items* en la escala deben formularse según expresen actitudes o juicios favorables o desfavorables con respecto a la actitud. *No se trata* de elegir *items* que expresen distintos puntos en el continuo.

f) Evite los *items* con los cuales todos o prácticamente nadie concuerda.

g) Los *items* deben ser formulados en lenguaje simple, claro y directo.

h) Solamente en casos excepcionales exceda de las 20 palabras cuando formule el *item*.

i) Un *item* debe contener sólo una frase lógica.

j) Los *items* que incluyan palabras como "todos" "siempre", "nadie", etc. deben omitirse.

k) De ser posible, los *items* deben ser formulados con frases simples, y no compuestas.

l) Use palabras que el entrevistado pueda comprender.

m) Evite las negaciones, particularmente las dobles negaciones.

n) Combine los *items* formulados positiva y negativamente en una proporción aproximada a 50 % — 50 %.

Un sistema que puede ser aplicado para eliminar muchos *items* dudosos o que dan demasiados hechos es el siguiente: cada miembro del grupo de investigación responde a los *items* asumiendo primero una actitud positiva hacia la variable y luego responde como si tuviese una actitud negativa. Si la respuesta en ambos casos se ubica en la misma categoría, el *item* no es apropiado para incluirse en la versión de los jueces.

Cada *item* es entonces un juicio o una sentencia a la cual el juez debe expresar *su* grado de acuerdo o desacuerdo. La graduación de acuerdos o desacuerdos varía en cantidad de alternativas que se le ofrece al sujeto; éstas pueden ser 3, 4, 5, 6 o 7 alternativas. En general la decisión sobre la cantidad de alternativas a ofrecer dependerá no tanto de las "preferencias personales" del investigador, sino del tipo de investigación, del tipo de pregunta, del tipo de distribución de la variable, etc. Ejemplificaremos a continuación algunos *items* con sus respectivas alternativas.

Las siguientes afirmaciones son opiniones con respecto a las cuales algunas personas están de acuerdo y otras en desacuerdo. Indique, por favor (marcando con una X en el paréntesis correspondiente), la alternativa *que más se asemeja a su opinión*.

1) Las mujeres no deberían meterse en política	( ) Muy de acuerdo ( ) De acuerdo ( ) Ni acuerdo ni desacuerdo ( ) En desacuerdo ( ) Muy en desacuerdo
2) Leyendo lo que se publica en los diarios y las informaciones de radio y TV, es posible tener una idea acertada de lo que ocurre en la situación política mexicana	( ) Muy de acuerdo ( ) De acuerdo ( ) En desacuerdo ( ) Muy en desacuerdo ( ) Ni acuerdo ni desacuerdo

[7] Edwards, A. L.: *Techniques of Attitude Scale Construction;* Appleton-Century-Crofts: Nueva York, 1967.

3) Los manifiestos, proclamas y solicitudes que publican en los diarios los partidos políticos no informan sobre sus verdaderos propósitos

( ) Totalmente de acuerdo
( ) Medianamente de acuerdo
( ) Escasamente de acuerdo
( ) Escasamente en desacuerdo
( ) Medianamente en desacuerdo
( ) Totalmente en desacuerdo

4) Vivo intensamente el presente sin pensar en el futuro.

( ) Verdadero
( ) Falso

Los ejemplos expresan *items* positivos y negativos; manifiestos y latentes y con distintas alternativas de respuesta. Los *items* son construidos a partir de juicios que expresan alguna relación postulada a nivel de la teoría sustantiva, y de observaciones empíricas de afirmaciones de grupos o sujetos que pertenecen a grupos o asociaciones que manifiestan la propiedad que se quiere medir. Los *items* así pueden ser extraídos de libros, publicaciones y artículos que tratan teóricamente sobre el objeto que se quiere medir. También puede el investigador acudir a análisis de contenidos sobre discursos o manifiestos de individuos y asociaciones (por ejemplo si se trata de medir radicalismo-conservadurismo, una fuente muy rica para la formulación de *items* son las declaraciones de grupos de interés: empresarios, grupos políticos de izquierda y de derecha, etc.). Otras estrategias para la construcción de *items* aparecen señaladas en la sección correspondiente a la escala Thurstone.

2) *La administración de los* items *a una muestra de jueces*

Una vez construidos los *items* (30 a 50) éstos van a ser distribuidos a una muestra de jueces (entre 50 a 100) los cuales deben ser seleccionados al azar de una población con características similares a aquella en la cual queremos aplicar la escala final. (Para los procedimientos de selección de la muestra ver el capítulo *Muestreo* en este manual.) Estos jueces responderán a cada uno de estos *items* su opinión. Las instrucciones a los jueces pueden ser dadas según el siguiente ejemplo:

EJEMPLO DE UNA VERSIÓN PRELIMINAR

El presente es un estudio de opiniones de estudiantes universitarios respecto a problemas actuales de la universidad.

A continuación se le presentará una serie de afirmaciones respecto a las cuales algunas personas están de acuerdo y otras en desacuerdo. Después de cada afirmación se presentarán cinco alternativas de respuestas posibles:

( ) Totalmente de acuerdo
( ) De acuerdo en general
( ) Ni de acuerdo ni en desacuerdo
( ) En desacuerdo en general
( ) Totalmente en desacuerdo.

Indique por favor —marcando con una cruz entre el paréntesis— *la alternativa que más se asemeje a su opinión*. Cuando no entienda alguna afirmación, ponga un signo de interrogación (?) al frente de ella. Trate de responder lo más rápido posible. Muchas gracias.

1. Las representaciones estudiantiles deberían participar en las decisiones sobre planes de estudio.
( ) Totalmente de acuerdo
( ) De acuerdo en general
( ) Ni de acuerdo ni en desacuerdo
( ) En desacuerdo en general
( ) Totalmente en desacuerdo

2. Las clases en las que el profesor tiene todo el control son las que mejor resultados producen en el aprendizaje.
( ) Totalmente de acuerdo
( ) De acuerdo en general
( ) Ni de acuerdo ni en desacuerdo
( ) En desacuerdo en general
( ) Totalmente en desacuerdo

3. El plan de estudios de la UNAM debe ser centralizado en la Secretaría de Educación Pública
( ) Totalmente de acuerdo
( ) De acuerdo en general
( ) Ni de acuerdo ni en desacuerdo
( ) En desacuerdo en general
( ) Totalmente en desacuerdo

4. El trabajo en grupo es más productivo que el trabajo individual.
( ) Totalmente de acuerdo
( ) De acuerdo en general
( ) Ni de acuerdo ni en desacuerdo
( ) En desacuerdo en general
( ) Totalmente en desacuerdo

5. Las carreras a las que el gobierno y la Universidad deberían prestarles más apoyo son aquellas centradas en las necesidades del país.
( ) Totalmente de acuerdo
( ) De acuerdo en general
( ) Ni de acuerdo ni en desacuerdo
( ) En desacuerdo en general
( ) Totalmente en desacuerdo

6. El alumno debe tener libertad en la elección de cuál es la mejor manera de controlar su rendimiento académico.
( ) Totalmente de acuerdo
( ) De acuerdo en general
( ) Ni de acuerdo ni en desacuerdo
( ) En desacuerdo en general
( ) Totalmente en desacuerdo

7. La única obligación de los alumnos es estudiar. Los planes de estudio son asunto de los profesores.
( ) Totalmente de acuerdo
( ) De acuerdo en general
( ) Ni de acuerdo ni en desacuerdo
( ) En desacuerdo en general
( ) Totalmente en desacuerdo

8. La mejor manera de juzgar a un estudiante es por su rendimiento académico.
( ) Totalmente de acuerdo
( ) De acuerdo en general
( ) Ni de acuerdo ni en desacuerdo

( ) En desacuerdo en general
( ) Totalmente en desacuerdo

9. Un trabajo hecho con consulta (en equipo o individualmente) permite una mejor evaluación de los conocimientos de los alumnos que una prueba hecha en la clase.
( ) Totalmente de acuerdo
( ) De acuerdo en general
( ) Ni de acuerdo ni en desacuerdo
( ) En desacuerdo en general
( ) Totalmente en desacuerdo

10. Es preferible que cada escuela universitaria tenga un programa fijo de estudios en vez de que, como sucede en otros países, el alumno pueda escoger con alguna libertad ciertas materias de su agrado.
( ) Totalmente de acuerdo
( ) De acuerdo en general
( ) Ni de acuerdo ni en desacuerdo
( ) En desacuerdo en general
( ) Totalmente en desacuerdo

11. Es preferible que los alumnos no hagan preguntas o intervenciones durante la exposición del profesor.
( ) Totalmente de acuerdo
( ) De acuerdo en general
( ) Ni de acuerdo ni en desacuerdo
( ) En desacuerdo en general
( ) Totalmente en desacuerdo

12. No conviene que los alumnos intervengan en la confección de los programas de estudio.
( ) Totalmente de acuerdo
( ) De acuerdo en general
( ) Ni de acuerdo ni en desacuerdo
( ) En desacuerdo en general
( ) Totalmente en desacuerdo

13. Las pruebas y exámenes deben limitarse exclusivamente a evaluar el grado de conocimiento de los alumnos respecto a la materia expuesta durante las horas de clase.
( ) Totalmente de acuerdo
( ) De acuerdo en general
( ) Ni de acuerdo ni en desacuerdo
( ) En desacuerdo en general
( ) Totalmente en desacuerdo

14. Es inconveniente que ocupen plazas en la universidad estudiantes que se verán impedidos de seguir sus estudios por falta de medios económicos.
( ) Totalmente de acuerdo
( ) De acuerdo en general
( ) Ni de acuerdo ni en desacuerdo
( ) En desacuerdo en general
( ) Totalmente en desacuerdo

15. Deben tener acceso a la cátedra universitaria sólo personalidades científicas de determinada orientación ideológica.
( ) Totalmente de acuerdo
( ) De acuerdo en general
( ) Ni de acuerdo ni en desacuerdo
( ) En desacuerdo en general
( ) Totalmente en desacuerdo

16. No habría por qué hacer esfuerzos en adecuar los horarios de clase para la gente que trabaja. O se trabaja o se estudia.
( ) Totalmente de acuerdo
( ) De acuerdo en general
( ) Ni de acuerdo ni en desacuerdo
( ) En desacuerdo en general
( ) Totalmente en desacuerdo

17. No deberían otorgarse becas a alumnos capaces, pero con recursos económicos suficientes, sino a aquellos de escasos recursos, aunque sean menos capaces.
( ) Totalmente de acuerdo
( ) De acuerdo en general
( ) Ni de acuerdo ni en desacuerdo
( ) En desacuerdo en general
( ) Totalmente en desacuerdo

18. Sólo a los organismos centrales de dirección de la enseñanza universitaria les está dada la facultad de decir a que ramos de la enseñanza han de conceder becas de estudios.
( ) Totalmente de acuerdo
( ) De acuerdo en general
( ) Ni de acuerdo ni en desacuerdo
( ) En desacuerdo en general
( ) Totalmente en desacuerdo

19. En el organismo universitario destinado a la distribución de becas de estudio entre las distintas facultades no deben participar estudiantes.
( ) Totalmente de acuerdo
( ) De acuerdo en general
( ) Ni de acuerdo ni en desacuerdo
( ) En desacuerdo en general
( ) Totalmente en desacuerdo

20. Para un mejor aprovechamiento de los recursos debe haber un organismo central que controle la concesión de becas de estudios.
( ) Totalmente de acuerdo
( ) De acuerdo en general
( ) Ni de acuerdo ni en desacuerdo
( ) En desacuerdo en general
( ) Totalmente en desacuerdo

21. A la hora de tomar un acuerdo importante sobre las evaluaciones docentes universitarias, las autoridades deben hacerlo sin tener en cuenta la opinión de los alumnos.
( ) Totalmente de acuerdo
( ) De acuerdo en general
( ) Ni de acuerdo ni en desacuerdo
( ) En desacuerdo en general
( ) Totalmente en desacuerdo

22. La universidad debería exigir de los egresados una retribución por los estudios recibidos, estableciendo un impuesto por el ejercicio profesional.
( ) Totalmente de acuerdo
( ) De acuerdo en general
( ) Ni de acuerdo ni en desacuerdo
( ) En desacuerdo en general
( ) Totalmente en desacuerdo

23. No compete a las escuelas universitarias fijar la cantidad de matrículas anuales para sus alumnos, sino a un organismo superior central.
( ) Totalmente de acuerdo
( ) De acuerdo en general
( ) Ni de acuerdo ni en desacuerdo
( ) En desacuerdo en general
( ) Totalmente en desacuerdo

24. La decisión respecto a la selección y otorgamiento a los estudiantes de material académico, debe ser tomada exclusivamente por los organismos centrales de la universidad.
( ) Totalmente de acuerdo
( ) De acuerdo en general
( ) Ni de acuerdo ni en desacuerdo
( ) En desacuerdo en general
( ) Totalmente en desacuerdo

25. La ampliación de las carreras universitarias existentes, debe responder so-
( ) Totalmente de acuerdo
( ) De acuerdo en general

lamente al número de los postulantes que se presenten en cada escuela.
( ) Ni de acuerdo ni en desacuerdo
( ) En desacuerdo en general
( ) Totalmente en desacuerdo

26. En las pruebas y exámenes escritos no debe haber personal universitario que vigile a los alumnos.
( ) Totalmente de acuerdo
( ) De acuerdo en general
( ) Ni de acuerdo ni en desacuerdo
( ) En desacuerdo en general
( ) Totalmente en desacuerdo

27. Los servicios de habitación y restaurante brindados por la universidad deben ser administrados sin ninguna participación del estudiantado.
( ) Totalmente de acuerdo
( ) De acuerdo en general
( ) Ni de acuerdo ni en desacuerdo
( ) En desacuerdo en general
( ) Totalmente en desacuerdo

28. No debería haber un organismo central de planificación universitaria.
( ) Totalmente de acuerdo
( ) De acuerdo en general
( ) Ni de acuerdo ni en desacuerdo
( ) En desacuerdo en general
( ) Totalmente en desacuerdo

29. Los alumnos no deben tener injerencia alguna en la labor del personal docente.
( ) Totalmente de acuerdo
( ) De acuerdo en general
( ) Ni de acuerdo ni en desacuerdo
( ) En desacuerdo en general
( ) Totalmente en desacuerdo

3) *Asignación de puntajes a los* items

Con este paso comienza efectivamente el análisis de la escala. Hay que clasificar a cada *item* según sea positivo o negativo, y luego ponderar las alternativas de respuesta. Nuevamente existen diferentes criterios para la adjudicación de las ponderaciones. Por ejemplo, los pesos para un *item positivo* pueden ser:

Pesos

4 ( ) Totalmente de acuerdo
3 ( ) De acuerdo en general
2 ( ) Ni de acuerdo ni en desacuerdo
1 ( ) En desacuerdo en general
0 ( ) Totalmente en desacuerdo

o la alternativa:

Pesos

2 ( ) Totalmente de acuerdo
1 ( ) De acuerdo en general

0 ( ) Ni de acuerdo ni en desacuerdo
−1 ( ) En desacuerdo en general
−2 ( ) Totalmente en desacuerdo

o cualquier otra serie de números.

Por lo general desaconsejamos la utilización de signos positivos y negativos en la adjudicación de los puntajes o de pesos a las alternativas de respuestas, ya que pueden crear la falsa impresión que la escala está midiendo a nivel intervalar; esto es, donde tendríamos puntajes finales en los que existe una posición 0, posiciones +1, +2..., +40; y posiciones −1, −2, −3..., −40. De hecho la escala mide a nivel ordinal y los valores de escala, simplemente implican posiciones de rango.

Para los *items* negativos, hay que recordar que la serie de números a adjudicar debe ser inversa. Por ejemplo, en un *item* negativo, la ponderación siguiendo la primera alternativa se haría:

Pesos

0 ( ) Totalmente de acuerdo
1 ( ) De acuerdo en general
2 ( ) Ni de acuerdo ni en desacuerdo
3 ( ) En desacuerdo en general
4 ( ) Totalmente en desacuerdo

Los *items* se ubican ya sea en forma positiva o negativa en relación a la variable con el fin de controlar los efectos del *"response-set"*, esto es, controlar las pautas de respuesta de aquellos respondentes que tienden a dar respuestas afirmativas o negativas de manera automática.

4) *Asignación de puntajes totales*

Este paso consiste simplemente en la adjudicación de los puntajes totales para cada individuo en la muestra de jueces. Esta suma resultará de la adición de los puntajes ponderados para cada *item*. En el caso de *items* con valores negativos, la suma es algebraica.

De comienzo estamos asumiendo que las personas con alto grado en la variable van a tener puntajes altos, mientras que las personas con una baja actitud manifestarán puntajes bajos. Si hemos presentado 30 *items* con un valor ponderado máximo de 4, y un mínimo de cero, la amplitud total de la dispersión de la variable a esperar sería entonces 120 (puntajes máximos de 120 y mínimo de 0 respectivamente).

5) *Análisis de los* items

Una vez computados los puntajes totales para todos los jueces, hay que ordenarlos de manera que el sujeto con el puntaje total más alto ocupe el

primer lugar, el segundo puntaje más elevado a continuación, etc., hasta llegar a la persona con el puntaje más bajo.

Una vez ordenados los sujetos, vamos a operar únicamente con los cuartiles superiores e inferiores es decir el 25 % de los sujetos con puntajes más elevados y el 25 % de los sujetos con puntajes más bajos. Del 50 % del centro no nos vamos a preocupar más. Formamos de esta manera un grupo *alto* y un grupo *bajo* con respecto a la variable y a los puntajes totales. Si tuviéramos 52 jueces, el grupo alto estará constituido por los 13 jueces con los puntajes más elevados y los 13 con puntajes más bajos.

Tomamos a estos 26 sujetos y los colocamos en una tabla en donde situamos las puntuaciones en cada *item* y el puntaje total para cada uno de los sujetos ordenados.

Hay que seleccionar ahora los *items* que discriminen mejor. Hay tres técnicas más en uso para la selección de los *items:* la del *cálculo del poder discriminatorio* de cada *item;* la de correlación *item-test;* y *el test de la mediana.*

El procedimiento para el cálculo del poder discriminatorio de un *item,* sigue la forma mencionada en el cuadro 3. Una vez separados el grupo alto y el grupo bajo, se calculan los promedios de cada *item* en cada uno de los grupos. Siguiendo el ejemplo de la figura 1, el promedio del *item* 1 es de 3.7 para el grupo alto (resultado de 48/13); y de 0.9 para el grupo bajo (resultado de 12/13).[8]

Una vez calculados los valores promedios para cada *item* en los grupos alto y bajo, procedemos a calcular el poder discriminativo de cada *item* según la fórmula:

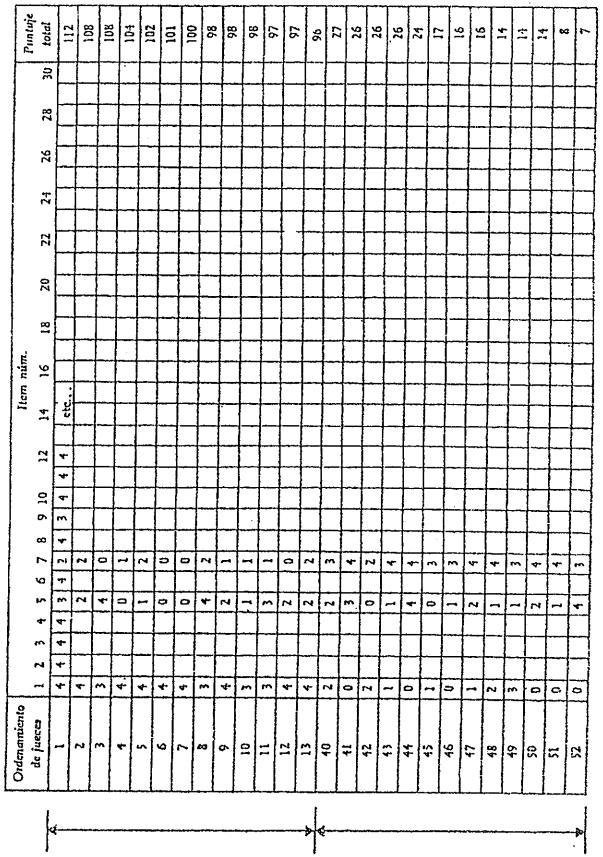
$$t = \frac{DM}{\sqrt{\frac{s^2_{M_1}}{N_1 - 1} + \frac{s^2_{M_2}}{N_2 - 1}}}$$

Donde:

t = *Test* t de Student
DM = Diferencia entre medidas $(M_1 - M_2)$
s^2 = variancias de muestra 1 y 2 respectivamente
N = cantidad de casos en cada una de las muestras

Para el cálculo del poder discriminativo del *item* conviene utilizar el siguiente cuadro, que es una continuación del cuadro 3:

[8] Para llegar al promedio sumamos la columna correspondiente al *item* 1 en los 13 sujetos con puntajes altos dividiéndolos por el total de casos: (4+4+3+4+4+4+4+3+4+3+3+4+4)/13. En forma idéntica, para el grupo bajo el cálculo es: (2+0+2+1+0+1+0+1+2+3+0+0+0)/13.



CUADRO 4. *Valores promedios para los grupos alto y bajo y diferencia de medias, para el cálculo del poder discriminatorio de* items *en la versión de los jueces en una escala Lickert*

	1	2	3	4	5	6	7	8	9	10	11	12
Promedio en el grupo alto (M_1)	3.7				1.8		1.1					
Promedio en el grupo bajo (M_2)	0.9				1.7		3.4					
Diferencia de medias ($M_1 - M_2$)	2.8				0.1		-2.3					

Una vez calculados los valores *t*, confrontamos con la tabla de distribución valores correspondientes, seleccionando aquellos *items* que realmente presenten diferencias significativas entre ambos grupos de contraste.

En los ejemplos señalados en la figura 1, la diferencia para el *item* 1 es significativa al nivel de 0.1 y la del *item* 7 al nivel de 0.5. El *item* 5 debe ser desechado porque no discrimina significativamente. Es importante notar asimismo, que en el caso del *item* 7 hemos colocado el signo del *item* originalmente mal ya que, como vemos, el grupo con valores bajos responde al estímulo en forma más positiva que los del grupo alto. Corresponde entonces cambiar los valores, manteniendo al *item*.

El método item-test para la selección de los *items* consiste en correlacionar el puntaje del *item* con el puntaje total del *test*. El coeficiente de correlación a utilizar es el coeficiente de correlación biserial, ya que aquí se trata de dos variables, una que podemos considerar intervalar, y a la otra una serie dicotomizada, siendo además la dicotomía forzada. El supuesto que tenemos que aceptar es que la distribución original es continua y normal. La fórmula para el cálculo de la correlación será:

$$r_b = \frac{M_p - M_q}{\sigma t} \cdot \frac{p \cdot q}{y}$$

Donde:

M_p y M_q = Medias parciales en el grupo alto y bajo, respectivamente.
σt = Desviación estándar total.
p y q = Proporción de casos en una y otra distribución.
y = Ordenada a la curva normal.

Cualquiera que sea la técnica utilizada para el análisis de los *items*, el objetivo es seleccionar aquellos que discriminen mejor (valores significativos de *t*, o valores altos de r_b).

De hecho hay una alternativa al uso de la prueba *t*, o del coeficiente de correlación biserial, que estrictamente se corresponde más con el tipo de medición de nivel ordinal, con el que opera la escala Lickert. En este caso la diferencia entre medianas, computada a través del "*test* de la mediana". Para ello hay que determinar primero el valor de la mediana de cada *item* para los grupos alto y bajo combinados, luego dicotomizamos los valores en una tabla de 2 × 2 para cada *item*, de la siguiente forma:

	Grupo alto	Grupo bajo	
Número de puntajes por debajo de la mediana combinada	A	B	A + B
Número de puntajes por encima de la mediana combinada	C	D	C + D
	A + C	B + D	

A esta tabla aplicamos ya sea x^2 o el *test* de Fisher, según sea la cantidad de casos (por lo menos $N = 40$ para aplicar x^2; con menos casos se recomienda el *test* de Fisher).

La fórmula para el cómputo de x^2 para este tipo de figura es:

$$x^2 = \frac{N\left[(AD - BC) - \frac{N}{2}\right]^2}{(A+B)\ (C+D)\ (A+C)\ (B+D)}$$

Seleccionamos por supuesto aquellos *items* cuyo x^2 da diferencias significativas.

Usando la misma figura, el *test* de Fisher se computa según la siguiente fórmula:

$$p = \frac{(A+B)!\ (C+D)!\ (A+C)!\ (B+D)!}{N!\ A!\ B!\ C!\ D!}$$

6) *La versión final de la escala*

La cantidad de *items* seleccionados de acuerdo a su poder discriminativo, constituirá la escala final a ser aplicada a sujetos o grupos como versión final. Los puntajes finales a adjudicar a los individuos, serán entonces el producto de la suma de los puntajes obtenidos en cada *item*, divididos entre el total de *items*.

B) *Comentarios finales*

Para calcular la *confiabilidad de la escala*, se puede utilizar la correlación entre mitades del *test* (*split-half reliability*): se correlacionan la suma de los

puntajes en los *items* impares con la suma de los puntajes de los *items* pares. Utilizamos p de Spearman, y luego la fórmula:

$$C = \frac{2p}{1+p}$$

A continuación presentamos las ventajas y desventajas de la escala Lickert, comparada con la Thurstone.

Ventajas. a) Permite la utilización de *items* que no se encuentran relacionados en forma manifiesta con la actitud que se desea estudiar (es decir, se pueden utilizar *items* con contenido latente). *b)* Es más rápida y fácil de construir. *c)* A mismo número de *items*, es más confiable. *d)* La cantidad de alternativas de respuesta permite una información más precisa de un sujeto en un *item* particular.

Desventajas. a) Por tratarse de una escala ordinal, no permite apreciar la distancia que hay entre pares de sujetos con respecto a la actitud. *b)* Con frecuencia dos puntajes iguales pueden ocultar pautas de respuestas diferentes de los individuos. *c)* No hay garantía de unidimensionalidad, consecuentemente pueden mezclarse distintas dimensiones, no estando seguro el investigador de cuál de ellas realmente se trata.

La escala Thurstone

Thurstone es quien provee la racionalidad —mediante su *ley de juicios comparativos*— para todo el aparato conceptual en la construcción de escalas para medir actitudes. Esta ley sostiene que para cada estímulo (e) dado, está asociado un proceso modal discriminal sobre un continuo psicológico. La distribución de todos estos procesos discriminacionales sigue la forma de la distribución normal, en la que todos los procesos discriminacionales producidos por el estímulo se distribuyen normalmente alrededor del proceso de discriminación modal, con una dispersión discriminal (s_i). Dado un conjunto n de estímulo, es posible ordenarlos en un continuo psicológico tomando como referencia el grado de atributo que ellos poseen.

A partir de estos principios, Thurstone propone su *escala de intervalos aparentemente iguales*, de tipo diferencial, en la que los *items* son seleccionados por una serie de técnicas que permiten escalonarlos de manera tal que expresen el continuo psicológico subyacente. La medición trata de establecerse al nivel intervalar. Es decir, una escala en la que sea posible afirmar que la distancia que separa a un sujeto que obtuvo una puntuación de 8.7 con respecto a otro sujeto que obtuvo 6.3, es igual a la distancia que separa a otro par de sujetos que obtuvieron puntuaciones de 3.6 y 1.2 respectivamente, y a cualquier distancia que sea igual a 2.4 puntos. (Sin embargo, como veremos más adelante, es discutido que la escala mida efectivamente a este nivel.)

El continuo psicológico en la escala de intervalos aparentemente iguales de Thurstone, se edifica sobre una serie de juicios de actitud distribuidos

en una escala de 11 puntos, en la que el punto 1 de la escala representa una actitud extrema (favorable o desfavorable), el punto 6 representa una actitud neutra (ni favorable ni desfavorable); y el punto 11 el otro extremo (favorable o desfavorable, según el extremo contrario a la actitud asumida en 1).

Los *items* en la escala Thurstone son construidos, diseñados y seleccionados de manera tal que permitan atribuir a los sujetos a los que se aplicará definitivamente la escala, un punto en un continuo. Así, esta escala es un poco más refinada que la escala Lickert, e implica una cantidad considerable de trabajo adicional.

A) *La construcción de una escala Thurstone*

Los procedimientos para la construcción de una escala de este tipo son: *1)* Se construye una serie de *items* (alrededor de 150). *2)* Se solicita a un grupo de jueces (más o menos 100), que ubiquen a los *items* en una escala de 11 puntos. *3)* Una vez evaluados los *items* por los jueces, se adjudica a los *items* valores de escala. *4)* Se seleccionan los *items* que representan el rango entero de la escala, rechazando los *items* ambiguos.

Detallamos cada uno de los pasos:

1) *La construcción de los* items *(versión de los jueces)*

Construya entre 100 y 200 *items* tomando en consideración los siguientes criterios (ver Edwards, A. L. para mayores detalles):

a) Evite los *items* que señalan al pasado en vez del presente.

b) Evite los *items* que dan demasiada información sobre hechos o los que fácilmente puedan ser interpretados como tales.

c) Evite los *items* ambiguos.

d) Evite los *items* irrelevantes con respecto a las actitudes que pretende medir.

e) Evite los *items* con los cuales todos o nadie concuerda.

f) Los *items* deben ser formulados en un lenguaje simple, claro y directo.

g) Sólo en casos excepcionales sobrepase las 20 palabras en un *item*.

h) Un *item* sólo debe contener una frase lógica.

i) Los *items* que incluyan palabras como "todos", "siempre", "nadie", "nunca", etc. serán percibidos de la misma manera y por ello deben omitirse.

j) De ser posible, los *items* deben ser formulados como frases simples y no compuestas.

k) Use sólo palabras que el entrevistado pueda comprender.

l) Evite las negaciones, especialmente las dobles negaciones.

m) Combine los *items* formulados positiva, neutral y negativamente en una proporción de 1/3, 1/3 y 1/3, distribuidos uniformemente sobre la variable.

Algunas maneras de formular *items* pueden ser: *i)* Extraerlos de libros, publicaciones y artículos que tratan sobre el objeto cuya actitud se quiere medir. Son importantes también las declaraciones y los discursos, a partir de los cuales uno pueda hacer una especie de análisis de contenido. Cuando se

sigue esta estrategia hay que tener en cuenta que un buen monto de reformulación va a ser necesario. *ii)* Concertar una discusión entre personas que representan distintos puntos de vista con respecto a la actitud. En este caso la grabación de la discusión facilitará la selección de frases adecuadas. Aquí también la reformulación será necesaria. *iii)* Formular uno mismo o en cooperación con otros investigadores, los enunciados ante los cuales se espera que la gente reaccionará en forma positiva, negativa o neutra.

En todo caso nunca es fácil llegar a la enunciación de 100-200 *items* sin incurrir en repeticiones o en formulaciones muy similares. Es importante, por cierto, que la distribución de los *items* a ser presentados a los jueces sea aproximadamente pareja, conteniendo un tercio de *items* positivos, negativos y neutros, a lo largo de un continuo.

2) *La administración a los jueces*

La lista de *items* (100 a 200), se distribuye a jueces (preferiblemente 200, con un mínimo de 50), los cuales van a ubicar a los *items* en una escala intervalo-subjetiva que va de 1 a 11 puntos.

Los jueces son seleccionados en función de su *conocimiento* sobre el problema que se quiera medir; y en la clasificación o evaluación de los *items* no importa la opinión personal del juez, sino su evaluación del punto en la escala continua de 1 a 11 en el cual él ubica al *item* (es decir, la determinación del peso que el *item* tiene en su opinión para la medición de la actitud).

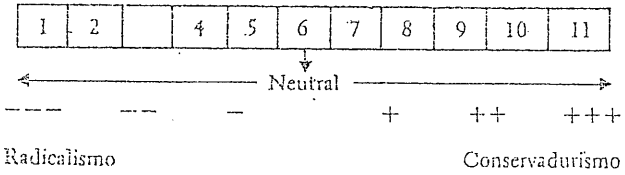
Las instrucciones a los jueces —que operan en forma independiente— pueden ser las siguientes:

EJEMPLO: ESCALA ACTITUD TIPO THURSTONE

Instrucciones

El Seminario de Recolección y Análisis de Datos del Centro de Estudios Sociológicos de El Colegio de México está realizando una serie de ejercicios sobre construcción de escalas para la medición de actitudes.

En el presente caso, tiene usted en sus manos una serie de afirmaciones para construir una escala de tipo Thurstone. La escala de intervalos aparentemente iguales de Thurstone parte de una serie de supuestos y técnicas en la cual los *items* son escalonados de manera tal que expresen un continuo subyacente. El continuo se edifica sobre una serie de juicios de actitud distribuidos en una escala de 11 puntos.



El punto 1 representa una actitud extrema (en nuestro caso: radicalismo), el punto 6 una actitud neutra (ni radical ni conservador) y el punto 11 el extremo conservador.

Para la construcción de esta escala, se requiere someter una cantidad de *items* a un número limitado de *jueces* antes de ser aplicada a una muestra de una determinada población.

Los jueces se eligen entre personas que tienen conocimientos especializados sobre la variable que se trata de medir, como ha sido el caso con usted.

En el presente caso, se trata de construir una escala que mida actitudes de radicalismo-conservadurismo. Quisiéramos pedirle que, en relación a esto, *no* vuelque usted *sus propias* opiniones acerca de las afirmaciones que aparecen a continuación, sino que usted exprese su juicio acerca de cuán radical, conservadora o neutra le parece cada una de las afirmaciones.

A la izquierda de cada afirmación hay un cuadrado en el cual usted debe colocar, de acuerdo a su criterio, el número que representa la frase en el continuo de 1 a 11, siendo:

1 el extremo "radical",

6 el punto medio o "neutral",

11 el extremo "conservador".

Si cree que la expresión se ubica *entre alguno de estos puntos*, utilice el número intermedio que mejor represente la posición de la frase. No trate de obtener el mismo número de *items*, o cualquier otra distribución espacial, en el continuo de 1 a 11.

Por favor, antes de empezar, lea una buena cantidad de expresiones para entender el carácter de los *items*.

Queremos insistir en que *no* se trata de dar una opinión personal de acuerdo o desacuerdo acerca de cada afirmación, sino solamente de estimar su lugar en una escala de 1 a 11 puntos.

[*Para el ejemplo presentamos solamente los primeros* 29 items]

☐ 1. El modelo desarrollista fortalece la desigualdad social.
☐ 2. Tratándose de programas políticos, "más vale malo conocido que bueno por conocer".
☐ 3. Todos los asuntos económicos de interés nacional deben estar a cargo exclusivo del Estado.
☐ 4. Los trabajadores deben tomar en sus manos la conducción del Estado; el Estado debe tomar en sus manos la conducción de la economía.
☐ 5. La desigualdad social ha existido siempre, y es necesaria para el desarrollo de la sociedad.
☐ 6. Los sindicatos deberían limitar sus actividades a las reivindicaciones económicas de sus representados.
☐ 7. La función del Estado es la de participar dinámicamente en el desarrollo económico, social y político del país.
☐ 8. La propiedad privada es un derecho natural del hombre y debe ser respetado y mantenido.

☐ 9. Las políticas sobre la distribución del ingreso son pura demagogia que sólo propician el enriquecimiento de los políticos.
☐ 10. La principal causa de la inflación es el anhelo de los empresarios de aumentar sus ganancias.
☐ 11. El Estado debería dejar absolutamente a criterio de los padres de familia, el tipo de educación que prefieren para sus hijos.
☐ 12. Todo grupo de interés debe tener igual representación ante el Estado.
☐ 13. Los trabajadores producen la riqueza; los patrones se la embolsan.
☐ 14. La política del control de la natalidad es una política al servicio del imperialismo.
☐ 15. Los que más pierden con la inflación son los empresarios.
☐ 16. La importancia creciente del sindicalismo en el país representa un peligro para la democracia.
☐ 17. Antes de la Reforma Agraria se producía más y mejor en el campo.
☐ 18. La mejor manera de resolver los problemas es encontrar el justo "término medio" y no caer en los extremos.
☐ 19. No se trata del sistema tal o cual; para que el país progrese hay que trabajar más, y punto.
☐ 20. El camino de México es el de una economía mixta: el gobierno como promotor, y la iniciativa privada como participante activa en el proceso de desarrollo.
☐ 21. El éxito se debe al esfuerzo personal.
☐ 22. La nacionalización de la industria minera sólo caería en el burocratismo y la mala administración estatal.
☐ 23. La mejor forma de representación ante el Estado es a través de grupos de interés, y no a través de partidos políticos.
☐ 24. La intervención del gobierno en la economía agrícola sólo ha traído desorden y caos.
☐ 25. El deficiente desarrollo agrícola del país se debe a la apatía y flojera de los campesinos.
☐ 26. La Revolución Mexicana será verdadera sólo cuando se realice una reforma agraria total.
☐ 27. La fuerza política de un grupo debe ser independiente del poder económico de sus miembros.
☐ 28. La actividad económica es privativa de los particulares y el Estado debe limitarse a coordinar tal actividad.
☐ 29. El marxismo es una doctrina exótica que no toma en cuenta nuestra idiosincrasia ni nuestra tradición.

3) *Asignación de valores de escala*

Vamos a estudiar ahora la distribución de las respuestas (de 1 a 11) en cada *item*, según las respuestas sobre ubicación del *item* dadas por los jueces. Podríamos calcular los valores promedios (media aritmética) y las desviaciones estándar (σ), para cada *item*. Sin embargo —y aquí de hecho se revela que la escala es más ordinal que intervalar— existe mayor exactitud y representa un método más rápido el calcular valores de mediana (*Mdn*) y de distancia intercuartil (Q_3-Q_1). En las siguientes páginas ilustramos con ejemplos de análisis gráficos. (Ver cuadros 5, 6 y 7.)

En el primer ejemplo (*Cuadro* 5) hemos obtenido una mediana de 2.3. La mediana indica el "valor" del *item* a lo largo de la variable, es decir que se trata en este caso de un *item* al lado positivo de la escala (muy dogmático). La distancia intercuartil es igual a 2,2. Esta cifra indica la "calidad" del *item*. Si menor la distancia intercuartil, mayor el grado de "calidad" del *item*, es decir, que la adjudicación del valor del *item* por parte de los jueces es similar. Los valores altos de distancia intercuartil, por el contrario, indican diferencias entre los jueces en cuanto a la apreciación sobre el valor adjudicado al *item*.

El ejemplo 2 (*Cuadro* 6) indica el caso ideal, en el que todos los jueces están de acuerdo en cuanto al valor que debe ser adjudicado al *item*.

El ejemplo 3 (*Cuadro* 7) representa un *item* de escasa utilidad, ya que como se ve los jueces le adjudicaron valores muy distintos, representando el valor mediano del *item* una adjudicación casi aleatoria de los jueces en cuanto al valor 1, 2, 3... u 11 que se le adjudique al *item* como expresando una posición en el continuo.

En los tres ejemplos presentados se han utilizado 56 jueces, y el método para calcular la mediana y los cuartiles ha sido proyectando líneas sobre la distribución gráfica. A partir de la distribución de frecuencias sobre los valores de *items* (1 a 11), resulta simple calcular la mediana y los cuartiles según las siguientes fórmulas:

$$Mdn = v + i\left(\frac{\frac{N}{2} - F_d}{F_p}\right)$$

Donde:

v_i = Límite exacto inferior del intervalo que contiene la mediana.
i = Amplitud del intervalo de clase (en nuestro caso igual a 1).
N = Número de casos.
F_b = Suma total de las frecuencias inferiores al intervalo que contiene la mediana.
F_p = Frecuencias del intervalo que contiene la mediana.

Para el cálculo de cuartiles:

$$Q_3 = v_s - i\left(\frac{\frac{N}{4} - F_a}{F_p}\right)$$

$$Q_1 = v_i + i\left(\frac{\frac{N}{4} - F_d}{F_p}\right)$$

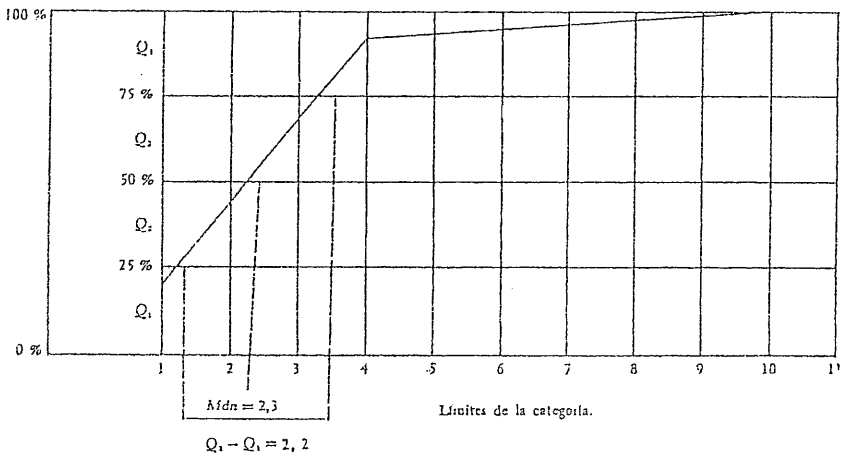
CUADRO 5. *Ejemplo 1: Escala de actitud* item *bueno tipo Thurstone mide Dogmatismo*

Item núm. 173

Caja núm.	1	2	3	4	5	6	7	8	9	10	11
Frecuencia	//// //// //	//// //// /	//// //// //	//// //// ///	////	/	//				/
Frecuencia en números	12	11	12	13	4	1	2				1
Frecuencia cumulativa	12	23	35	48	52	53	55				56
Porcentaje cumulativo	21.4	41.1	62.5	85.7	92.5	94.6	98.3				100
	1	2	3	4	5	6	7	8	9	10	11

56 jueces

Representación gráfica y cálculo de Mdn y Q



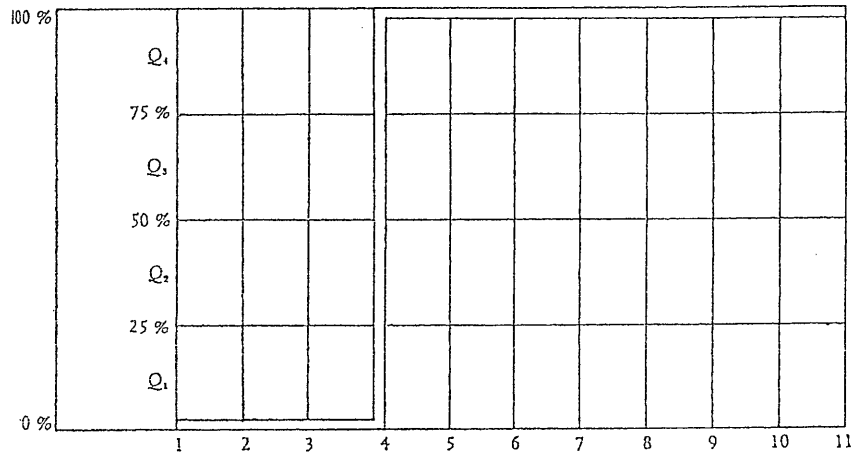
CUADRO 6. *Ejemplo 2. Escala de actitudes tipo Thurstone. Item ideal.*

Item núm. X

Caja núm.	1	2	3	4	5	6	7	8	9	10	11
Frecuencia											
Frecuencia en números	0	0	0	56	0	0	0	0	0	0	0
Frecuencia comulativa	0	0	0	56	56	56	56	56	56	56	56
Porcentaje comulativo	0	0	0	100	100	100	100	100	100	100	100
	1	2	3	4	5	6	7	8	9	10	11

56 jueces

Representación gráfica y cálculo de Mdn y A

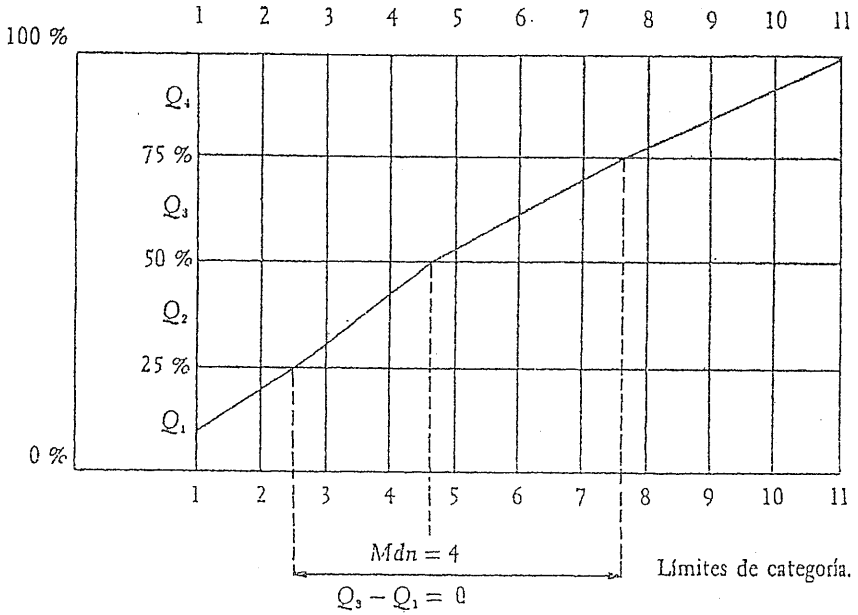


Ítem núm. 172

Caja núm.	1	2	3	4	5	6	7	8	9	10	11
Frecuencia	𝍸	𝍸 |	𝍸 |||	𝍸 |	𝍸 |	𝍸	𝍸 |	𝍸	𝍸	||	𝍸
Frecuencia en números	3	7	8	6	6	4	6	4	5	2	5
Frecuencia comulativa	3	10	18	24	30	34	40	44	49	51	5
Porcentaje comulativo	5.4	17.8	32.2	42.9	53.6	60.8	71.2	78.5	87.5	91 0	100
	1	2	3	4	5	6	7	8	9	10	11

56 jueces

Representación gráfica y cálculo de Mdn y A.



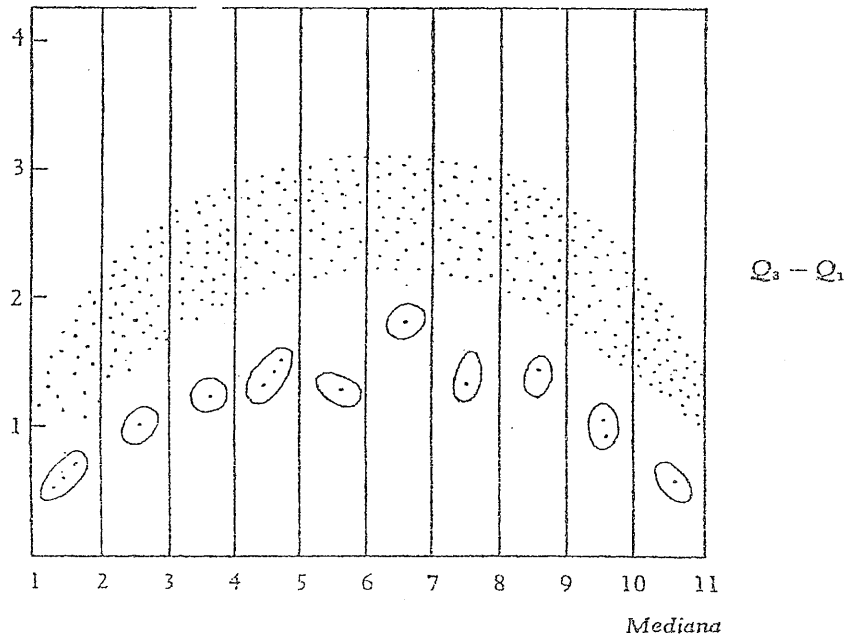
Donde:

v_s = Límite exacto superior del intervalo que contiene el cuartil.
v_i = Límite exacto inferior del intervalo que contiene el cuartil.
i = Amplitud del intervalo de clase.
F_a = Frecuencias por encima del intervalo que contiene el cuartil.
F_d = Suma total de frecuencias por debajo del intervalo que contiene el cuartil.
F_p = Frecuencias en el intervalo que contiene el cuartil.

4) *La selección de los* items

La selección final de los *items* presentados a los jueces se realiza en base a los valores de la mediana, y de amplitud intercuartil. La mediana se utiliza para ubicar el peso del *item* en la escala, eligiendo los *items* que se hallen repartidos uniformemente a lo largo de la misma. Necesitamos aproximadamente 2 *items* para cada intervalo de la escala (1-2; 2-3; 3-4 ... 9-10; 10-11).

FIGURA 3.



La distancia intercuartil es utilizada para decidir cuáles son los mejores *items* dentro de cada intervalo. La Figura 3 indica una distribución de *items* en la cual en el eje horizontal figuran los valores de mediana; y en el eje vertical los valores de distancia intercuartil. Los *items* seleccionados para la escala final están marcados con círculos y son obviamente aquellos que representando cada uno de los intervalos en la escala, presentan distancias intercuartiles mínimas.

Normalmente las distancias intercuartiles más grandes se presentan en el centro de la escala. Ocurre muchas veces que la selección de *items* para los valores 5, 6, 7 y 8 se complica en la medida que los jueces tienden a dispersar sus evaluaciones fuertemente en el centro; es decir, esto se va a reflejar en distancias intercuartiles bastante grandes. Si es que no se obtienen suficientes *items* en la primera versión de los jueces, el investigador deberá construir más *items* para esta área, que deben ser juzgados nuevamente por los mismos jueces. Conviene por supuesto anticiparse a este tipo de problemas, teniendo especial cuidado cuando se construyen los *items* iniciales.

La versión final de la escala

Está constituida por los 15 o 25 *items* seleccionados como confiables. Los *items* son presentados a los sujetos solamente con dos alternativas de respuestas: acuerdo-desacuerdo.

La forma de presentación de la escala es diferente de aquella presentada a los jueces. En un cuestionario, por ejemplo, se eliminan las "cajas", solicitando al sujeto que responda únicamente si está de acuerdo o en desacuerdo con cada uno de los juicios.

A manera de ejemplo, presentamos algunos *items* que han sido contestados por un sujeto hipotético.

1. El éxito se debe al esfuerzo personal.	☒ De acuerdo ☐ En desacuerdo
2. El futuro nos deparará mejores condiciones de vida.	☐ De acuerdo ☒ En desacuerdo
3. El gobierno debería tomar todas las decisiones.	☐ De acuerdo ☒ En desacuerdo
4. Casi nadie considera el trabajo que yo hago.	☐ De acuerdo ☒ En desacuerdo
5. Es bueno que la Iglesia se modernice.	☐ De acuerdo ☒ En desacuerdo
6. Yo no podría ser bueno de carácter.	☒ De acuerdo ☐ En desacuerdo

Etcétera.

5) *La adjudicación de puntajes a los sujetos*

El investigador tiene que registrar en un código los valores de mediana adjudicados por los jueces a los *items* seleccionados para integrar la versión final de la escala. Los valores intercuartiles no se utilizan para el cómputo de los puntajes finales de los sujetos. El puntaje final de un sujeto será entonces simplemente el promedio de los valores de escala de los *items* respondidos en forma afirmativa o "de acuerdo".

Supongamos que nuestro sujeto hipotético haya respondido "de acuerdo" únicamente a los *items* 1 y 6; es decir que en el resto de los *items* su respuesta ha sido "en desacuerdo". Supongamos entonces que los valores de escala dados por los jueces hayan sido:

Item	*Peso del* item *en la variable*
1	1.2
2	4.6
3	7.8
4	3.7
5	10.8
6	1.6
7	9.8
8	5.6
etc.	etc.

Consecuentemente, el puntaje correspondiente a nuestro sujeto será:

$$\frac{1.2 + 1.6}{2} = 1.4$$

Los valores en la escala utilizados como una variable en el análisis de alguna investigación, pueden también ser utilizados de varias formas. Se pue-

Peso en la escala Thurstone		*Núm. de individuos*	*Frecuencias acumulativas*	
1		34	34	
2		54	88	
3		42	130	
4		39	169	
5		15	184	grupo alto
6		5	189	
7		52	241	
8	Dicotomización	84	325	
9		82		grupo bajo
10		72		
11		121		
		600		

den tomar los distintos valores en los sujetos y establecer correlaciones entre la actitud y alguna otra variable, o bien dicotomizar las respuestas de la siguiente manera:

Supongamos que la distribución en nuestra variable entre 600 individuos en una encuesta es como aparece en la página anterior (supongamos que todos han contestado la pregunta).

Lo que hicimos con la distribución fue lo siguiente:

a) Contabilizamos el número de sujetos que obtuvieron los diferentes puntajes en la escala; aquí particularmente ubicamos a los sujetos que obtuvieron, por ejemplo, 4.6 en el intervalo 5. Nótese que entonces, la amplitud de cada intervalo de "peso en la escala", para el caso del intervalo 5, va de 4.5 a 5.49.

b) Calculamos las frecuencias acumuladas.

c) Si buscamos dicotomizar la variable, el estadístico a utilizar sería la mediana, que en nuestro caso tiene un valor igual a 8.2,. es decir, cae en la categoría 7.5 a 8.49.

d) Procedemos entonces a la dicotomización y obtenemos 2 grupos, a los cuales les llamaremos grupo *alto* y grupo *bajo*.

e) Podemos cruzar entonces nuestra variable con alguna otra, tal como figura en el ejemplo,

Variable *Y* (otra variable cualquiera)

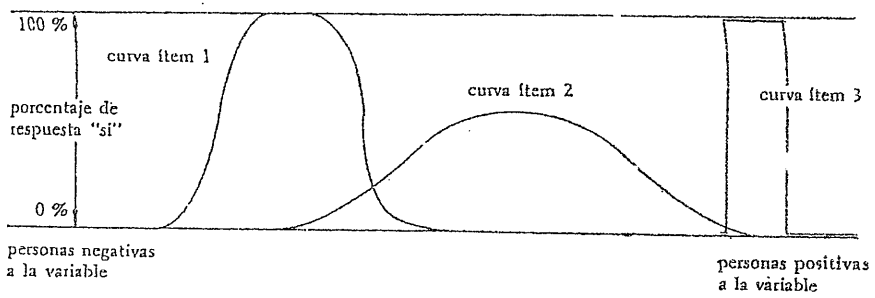
		Alto	Mediano	Bajo	
Variable X (Thurstone)	Grupo alto	15	130	180	325
	Grupo bajo	221	52	2	275
		236	182	182	600

C) *Comentarios*

1) Items *o series de* items *acumulativos o diferenciados*

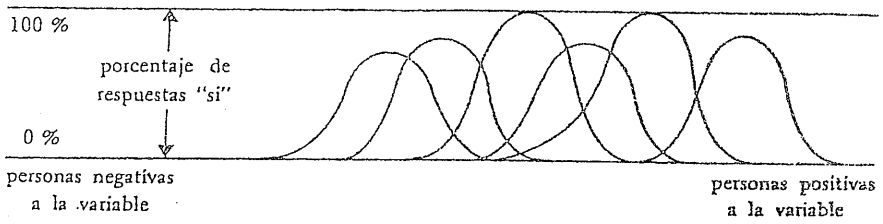
Los *items* de la escala Thurstone son *items* diferenciadores. Presentamos algunos ejemplos de diferentes *items* en una escala Thurstone. (Ver en la página siguiente.)

El eje horizontal representa todas las personas investigadas ordenadas en tal forma que las que detentan una actitud más negativa con respecto a la variable se ubican a la izquierda y las más positivas a la derecha. En el eje vertical está representado el porcentaje de respuestas "de acuerdo" en un *item*.



Item 1: es un *item* que recibe respuestas "No" de las personas que tienen una fuerte posición ya sea positiva o negativa. Un grupo de personas que tengan una determinada posición a lo largo de la escala, responderán este *item* con un "Sí". En el caso del *item* 2, la única diferencia es que solamente el 50 % de las personas responderá "Sí" al *item* en el lugar de la escala donde la probabilidad de alta tasa de respuestas "Sí" es más grande. El *item* 3 es un *item* donde el 100 % de los respondentes con ciertas posiciones en la escala responderán "Sí" y todos los demás "No". Ejemplo: "¿Está usted entre los 178 y los 179 cm de estatura?"

La escala Thurstone en su totalidad forma una serie de *items* diferenciadores. La siguiente figura puede dar una idea de la estructura de una escala Thurstone. Hemos reducido el número de *items* a 6 para facilitar la lectura de la figura.



2) *La variación en el instrumento, los sujetos o ambos*

En la escala de Thurstone, solamente el estímulo aislado recibe valor en la escala. El conjunto de tareas para los jueces es escoger los *items* dentro de los intervalos de igualdad-apariencia (11 cajas) tratando de disminuir cualquiera variación debido a su *propia* posición con respecto a la actitud. Es decir que en el caso de la escala Thurstone tenemos el enfoque que corresponde a la variación centrado en el instrumento. La escala Thurstone entonces pertenece al *test* Tipo B de Coombs.

3) *El problema con las dimensiones*

Una desventaja de la escala de Thurstone es la dificultad de controlar las dimensiones en el variable. Supongamos que una persona ha contestado "de acuerdo" a los siguientes *items* de una escala Thurstone de 30 *items*:

Item	*Valor*
4	5.9
9	6.2
11	6.9
19	7.2
28	7.8
	34.0

Su promedio es entonces $\frac{34.0}{5} = 6.8$

Pudo haber recibido esa suma particular mediante una serie de operaciones. Los cinco *items* con los que concordó pudieron haber tenido asignados los números 4.1; 4.2; 8.7; 8.5; 8.9 o por ejemplo: 1.9; 2.2; 7.4; 10.7 y 10.8. En ambos casos el promedio es 6.8. Las posibilidades aumentan todavía más si consideramos que, de hecho, el individuo podía haber concordado con más o menos de 5 *items*. ¿Cuál podría ser la razón de esto? Como simplificación gruesa del problema podemos pensar que la serie 4.1; 4.2; 8.7; 8.5 y 8.9 indica dos dimensiones, una de las cuales aparece alrededor de 4 y la otra alrededor de 9. Aparentemente los dos primeros *items* tienen para el sujeto un significado distinto que los tres siguientes. Para controlar esta desventaja de la escala, Thurstone y Chave (1929) proveyeron a la escala de un criterio burdo basado en las respuestas: el llamado *index of similarity* (índice de similitud).

D) *Ventajas y desventajas de la escala Thurstone*

Ventajas. a) La principal ventaja de esta técnica es que permite hacer una distribución de un grupo dado, a lo largo de la actitud que se desea investigar, precisamente porque los *items* fueron diseñados y seleccionados a los efectos de cubrir el continuo. *b)* Supone una medida más refinada que la escala Lickert, ya que el puntaje de los *items* se deriva de una ponderación basada en el juicio de jueces expertos o al menos informados. *c)* En la medida que la escala final contiene más *items* que la de Lickert, es más confiable que ésta. *d)* Si es tratada como escala intervalar, permite comparar puntuaciones y cambios de actitud en los sujetos.

Desventajas. a) Su elaboración es larga y compleja. *b)* Pese a su tratamiento como escala intervalar, su verdadera naturaleza en términos de nivel de medición corresponde al nivel ordinal. *c)* Fácilmente se introducen otras dimensiones distintas de las que se quiere medir. *d)* Discrimina poco en los extremos de la distribución. *e)* Los *items* neutrales carecen de significado mayor, y a menudo se ubican en esta posición *items* que no se refieren a la dimensión tratada. *f)* Distintas configuraciones de respuesta resultan en el mismo puntaje final. *g)* Los jueces pueden introducir sesgos difíciles de detectar.

ESCALAS PARA LA MEDICIÓN DE ACTITUDES

LA ESCALA GUTTMAN

Una de las desventajas mayores en las dos escalas que examinamos hasta ahora —Lickert y Thurstone— era que ninguna de ellas garantizaba que el instrumento mida una dimensión única.

La escala Guttman, conocida como método del escalograma o análisis de escalograma, soluciona el problema de la unidimensionalidad. Su objetivo es el de definir lo más claramente posible qué es lo que está midiendo la escala, entendido esto como un problema de unidimensionalidad. Por el tipo especial de tratamiento al que se somete a la escala se busca la eliminación de factores extraños a la característica o dimensión que se pretende medir.

La escala Guttman es de tipo acumulativo, ya que la respuesta positiva a un *item* supone que los *items* anteriores también han sido respondidos en forma positiva. Se busca pues una coherencia en las pautas de respuesta de los sujetos, y esa coherencia es garantizada por medio de un *coeficiente de reproductividad*. El tamaño del coeficiente (valor máximo 1.00) señala el grado por el cual la escala es acumulativamente perfecta o casi perfecta. En una escala cuya reproductividad es perfecta, las respuestas de los sujetos a todos los *items* pueden ser reproducidas por el solo conocimiento de su posición de rango.

Veremos más adelante que además de la reproductividad, hay que tomar en cuenta otros factores, tales como el alcance de la distribución marginal, la pauta de los errores, el número de *items* en la escala y el número de categorías de respuesta.

El método de escalograma de Guttman combina aspectos de construcción utilizados en las escalas Lickert y Thurstone, a más de los distintos cálculos de los coeficientes mencionados en el parágrafo anterior, en razón de que utiliza 2 técnicas: *a)* Siguiendo los procedimientos de la escala Lickert; y *b)* La técnica de la escala discriminatoria de Edwards y Kilpatrick. Vamos a desarrollar primero los procedimientos siguiendo las técnicas de Lickert y luego utilizaremos la técnica de Edwards y Kilpatrick.

A) *La construcción de un escalograma Guttman*

Los pasos a dar en la construcción de una escala de este tipo son: *1)* Se construye una serie de *items* relevantes a la actitud que se quiere medir. *2)* Se administran los *items* a una muestra de sujetos que van a actuar como jueces. *3)* Se asignan puntajes a los *items* según la dirección positiva o negativa del *item*. *4)* Análisis de *items* para la formación de series acumulativas. *5)* En base a los *items* seleccionados se construye la escala final.

Para los pasos 1, 2 y 3 se siguen los procedimientos señalados en la escala

Lickert para la construcción de *items*, la administración a los jueces y la asignación de puntajes (ver escala Lickert).

4) *El análisis de los* items

Una vez aplicados los *items* a los jueces, procedemos al análisis de los *items* en su conjunto. La idea es formar una serie acumulativa de *items*. Para ello, vamos a diseñar un escalograma.

a) Primero computamos el puntaje total para cada uno de los jueces (es decir sumamos los valores obtenidos en cada uno de los *items*).

b) Ordenamos a los jueces según el puntaje total, desde el puntaje más alto al puntaje más bajo.

Si nuestra serie de *items* fuese perfecta, todas las celdas cruzadas (*crossed cells*) en el escalograma estarían en una posición sobre una diagonal que corre desde el ángulo superior izquierdo hasta el ángulo inferior derecho. Mientras más alto sea el número de desviaciones a esta diagonal, más baja será la reproductividad y menos idéntica a una serie de *ítems* acumulativos será nuestra serie.

Antes de entrar a un análisis en detalle del Cuadro 8, vamos a mostrar un escalograma más simple con el fin de clarificar la idea de *acumulación* y de *error*.

Supongamos que 20 jueces hayan respondido a 6 *items* en términos de "acuerdo" o "desacuerdo". Ordenamos a los jueces, registrando en el Cuadro 8 los "acuerdos" con una x, y los "desacuerdos" con una O. La distribución se muestra en el Cuadro 8.

El Cuadro 8 muestra una serie de *items* escalonados. El *item* 4 ocupa la primera posición en la escala en razón de que se dieron en él 3 respuestas positivas (de acuerdo); el *item* 2 ocupa el segundo lugar ya que se dieron 6 respuestas positivas, y así sucesivamente hasta llegar al último *item*, el número 3, con el que 16 de los 20 jueces acordaron.

En términos de escalograma, las respuestas redondeadas con un círculo son "errores", esto es, respuestas que caen fuera de la pauta general del escalograma señalado con la línea segmentada.

El cuadro analítico para el Cuadro 8, sería:

Número de preguntas: 6
Número de jueces: 20
Cantidad total de respuestas: $6 \times 20 = 120$
Cantidad de errores: 7

$$\text{Reproducibilidad} = 1 - \left(\frac{\text{Cantidad de errores}}{\text{Cantidad total de respuestas}} \right) = 0.94$$

Compliquemos un poco el ejemplo, utilizando 5 categorías o alternativas de cada *item*.

Supongamos que 30 jueces han juzgado 6 *items* en una escala de 5 puntos:

4— Muy de acuerdo
3— De acuerdo
2— Indeciso
1— En desacuerdo
0— Muy en desacuerdo

CUADRO 8. *Análisis de escalograma. Respuestas de 20 jueces a* 6 items *en términos de acuerdo-desacuerdo*

Rango del Juez							*Puntaje*
1	X	X	X	X	X	X	6
2	X	X	X	X	X	X	6
3	X	X	X	X	X	X	6
4	0	X	X	X	X	X	5
5	0	X	(0)	X	X	X	4
6	0	0	X	X	X	X	4
7	0	0	X	X	(0)	X	3
8	0	0	X	X	X	(0)	3
9	0	0	0	X	X	X	3
10	0	0	0	X	X	X	3
11	0	0	0	X	X	X	3
12	0	0	(X)	0	X	X	3
13	0	(X)	0	0	X	X	3
14	0	0	0	0	X	X	2
15	0	0	0	0	0	X	1
16	0	0	0	0	0	X	1
17	0	0	0	0	(X)	(0)	1
18	0	0	0	0	0	X	1
19	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0
	3	6	8	11	14	16	

Es decir, que teóricamente los puntajes van a variar desde un máximo de 24 puntos a un mínimo de 0.

El Cuadro 9 señala la ordenación de los jueces, y el puntaje obtenido en cada uno de los *items*. Considerando la dispersión sobre todas las alternativas de la escala, vemos que la probabilidad de obtener una escala acumulativamente perfecta disminuye, en la medida que se trata ahora de que haya escalabilidad entre los *items* y en el interior de un *item*.

Si examinamos el *item* 1 encontramos que hay 8 personas con respuesta 4 ("muy de acuerdo"); 10 personas con respuesta 3 ("de acuerdo"); 4 personas con respuesta 2 ("indeciso"); 4 personas con respuesta 1 ("en desacuerdo"),

CUADRO 9. *Análisis de escalograma. Ordenación de puntajes y valores de respuesta en 6 items tipo escala Lickert (30 jueces)*

Puntaje	*Item 1*					*Item 2*					*Item 3*					*Item 4*					*Item 5*					*Item 6*				
	4	*3*	*2*	*1*	*0*	*4*	*3*	*2*	*1*	*0*	*4*	*3*	*2*	*1*	*0*	*4*	*3*	*2*	*1*	*0*	*4*	*3*	*2*	*1*	*0*	*4*	*3*	*2*	*1*	*0*
24	X					X					X					X					X					X				
23	X					X						X				X						X					X			
21	X					X					X						X					X				X				
20		X					X				X						X					X				X				
19	X						X				X					X							X					X		
19	X						X				X								X			X				X				
19		X				X						X						X			X						X			
18	X						X				X								X			X					X			
18		X					X					X				X						X					X			
18		X					X					X				X						X					X			
18	X						X					X				X							X				X			
17		X				X							X					X				X					X			
16		X					X				X								X					X		X				
15		X								X	X							X				X					X			
15				X		X						X								X		X				X				
13		X							X			X							X					X		X				
12		X						X					X							X			X				X			
12		X							X						X				X	X		X					X			
11	X								X						X						X							X		
10			X					X					X							X				X			X			
8				X						X			X						X						X	X				
8				X					X				X							X			X				X			
8					X				X				X							X		X					X			
8			X							X					X				X					X			X			
7			X					X					X							X				X					X	
6					X			X				X								X				X						X
4			X					X							X					X					X					X
4				X					X				X							X			X							X
3				X						X					X					X				X					X	
2					X					X					X					X				X					X	

y 4 personas con respuesta 0 ("muy en desacuerdo"). De haber escalabilidad perfecta en el *item*, deberíamos esperar que todos los jueces con respuesta 4, se coloquen en un orden de rango por encima de aquellos que dieron respuesta 3. No ocurre así, ya que solamente 3 de ellos están claramente por encima, mientras que hay sujetos con puntaje 3 e incluso un sujeto con puntaje 1 que aparecen en un orden de rango superior. Si observamos los sujetos con puntuación 2 vemos que éstos están perfectamente ordenados en relación a los que tienen puntaje 3 y 4. En los sujetos con puntaje 1 vemos que existen dos sujetos mal ubicados (dos errores) en relación al ordenamiento escalable. De los 4 sujetos que tienen puntaje 0 en el *item*, 3 están fuera de posición.

Podríamos hacer un escalograma con el *item* 1, de la misma manera como lo hicimos en el Cuadro 1 y contar el número de errores.

Considerando la dispersión sobre todas las alternativas de la escala, la probabilidad de obtener una serie acumulativa de *items* en el ejemplo del Cuadro 9 disminuye, en relación con el ejemplo del Cuadro 8. A través de la unificación de algunas categorías de respuesta de distintas maneras, podríamos tratar de aumentar la reproductividad de la escala.

Las técnicas que se utilizan para este propósito son diferentes. Se puede recurrir directamente a máquinas electrónicas para procesamiento de datos, a alguna técnica de "tablas de escalograma" *(scalogram board techniques)*, o un procedimiento más complicado y que implica pérdida de tiempo como el de "ensayo y error".

Volviendo al ejemplo original del Cuadro 9, tenemos errores o muchos sujetos fuera de posición dondequiera que pongamos la diagonal o los cortes. Se tratará entonces de combinar las respuestas de manera que podamos representar un nuevo *continuum* en el cual, por ejemplo, se combinen respuestas "muy de acuerdo" con respuestas "de acuerdo", o cualquier combinación que indique escalabilidad. Esto significa que el criterio por el que nos guiamos para la combinación de respuestas es empírico, esperando que si efectivamente el *item* discrimina, ordenará las pautas de discriminación con una combinación tal que minimice el error. Cuando el *item* no discrimina, es decir, cuando cualquier combinación produce un número grande de errores (y grande significa simplemente que el número de aciertos es menor que el número de errores, o que no existe una desproporción significativa entre aciertos y errores que nos dará como resultado un coeficiente de reproducción satisfactorio), el *item* no es escalable y se elimina.

En el *item* 1 podemos dejar las categorías 4 y 3 separadas, y combinar las categorías de respuesta 2, 1 y 0. En el *item* 2 combinamos las categorías de respuesta 4 y 3, por un lado, y 2, 1 y 0 por el otro, etcétera.

Supóngase que las combinaciones propuestas para cada *item* y sus respectivos pesos son ahora:

Item	*Combinaciones*	*Nuevos pesos*
1	4 3 2,1,0	2 1 0
2	4,3 2,1,0	2 0
3	4,3 2,1,0	2 0
4	4,3,2 1,0	2 0
5	4,3 2,1,0	2 0
6	4,3,2 1,0	2 0

La distribución de los puntajes en los *items* adquiere ahora la siguiente forma:

CUADRO 10

	Item 1			*Item 2*		*Item 3*		*Item 4*		*Item 5*		*Item 6*	
puntajes	2	1	0	2	0	2	0	2	0	2	0	2	0
12	X			X		X		X		X		X	
12	X			X		X		X		X		X	
12	X			X		X		X		X		X	
11		X		X		X		X		X		X	
11		X		X		X		X		X		X	
11		X		X		X		X		X		X	
11		X		X		X		X		X		X	
10				X		X		X		X		X	
10	X			X		X			X	X	X	X	
10	X			X		X			X	X		X	
10	X			X		X		X			X	X	
9		X		X			X	X			X	X	
9		X			X	X		X		X		X	
8			X	X		X			X	X		X	
7		X		X		X			X		X	X	
6	X				X		X		X	X		X	
5		X			X	X			X		X	X	
5		X			X		X		X	X		X	
5		X			X		X		X	X		X	
4			X		X		X	X			X	X	
4			X		X		X		X	X		X	
4			X		X		X	X			X	X	
2			X		X		X		X		X	X	
2			X		X		X	X			X	X	
2			X		X		X		X		X	X	X
0			X		X		X		X		X	X	X
0			X		X		X		X		X	X	X
0			X		X		X		X		X	X	X
0			X		X		X		X		X	X	X

Obsérvese que con la nueva clasificación de los *items*, el *item* 6 no contiene ahora ningún error, siendo a su vez el que menos discrimina entre los puntajes extremos. Los *items* 2 y 3 escalan bastante bien y con poco error. El

item 4 tiene algunos errores mientras que el 5, a pesar de la dicotomización, contiene más errores que los demás. La distribución de los puntajes en el *item* 1 indica la necesidad de unir las nuevas categorías 2 y 1 con el fin de

CUADRO 11

	Items				
Puntajes	4	2	3	1	6
5	X	X	X	X	X
5	X	X	X	X	X
5	X	X	X	X	X
5	X	X	X	X	X
5	X	X	X	X	X
5	X	X	X	X	X
5	X	X	X	X	X
5	X	X	X	X	X
5	X	X	X	X	X
4	0	X	X	X	X
4	0	X	X	X	X
4	(X)	X	(0)	X	X
4	(X)	(0)	X	X	X
4	0	X	X	X	X
3	0	X	X	(0)	X
3	0	0	X	X	X
2	0	0	0	X	X
2	0	0	0	X	X
2	0	0	0	X	X
2	0	0	0	X	X
2	(X)	0	0	0	X
1	0	0	0	0	X
1	0	0	0	0	X
1	0	0	0	0	X
1	(X)	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0

reducir aún más el error. Tal como está el cuadro, su coeficiente de reproductibilidad será igual a:

$$r_p = 1 - \frac{\text{Total de errores}}{\text{Total de respuestas}} = 1 - \frac{38}{390} = .913$$

Valor satisfactorio, sin embargo, si unimos las categorías 1 y 2 del *item* 1, y eliminamos el *item* 5; detenemos entonces el escalograma de la página anterior, en el cual los *items* aparecen ordenados por grado de dificultad.

Puede verse ahora que los errores han sido reducidos en forma considerable, aunque a costa de reducir la discriminación al interior de los *ítems,* además de eliminar el *item* originalmente designado con el número 5.

El lector notará que con los nuevos pesos asignados a los *items* se produce un reordenamiento en el rango de los jueces como consecuencia de la alteración en los puntajes totales.

Las líneas segmentadas representan los puntos de cortes.

B) *La determinación de los errores*

Hay dos métodos que tienden a dar resultados diferentes: la técnica de Cornell y la técnica de Goodenough.

Comencemos con la técnica de Cornell, la cual se realiza estableciendo puntos de separación en el orden de rango de los jueces, puntos que los separan en distintas categorías, tales como se definirían de ser la escala perfecta. Ésta es la técnica utilizada cuando señalábamos puntos de separación por medio de líneas segmentadas. Si observamos el Cuadro 10, vemos que el *item* 1 necesita dos puntos de separación: el primero cae entre la última persona con puntaje 10 y la primera persona con puntaje 9. Todos los sujetos o jueces por encima de este punto de separación deberían tener puntaje 2 en el *item*. En términos de errores encontramos pues 5 errores (cuatro por encima y uno por debajo del punto de separación). El siguiente punto de separación lo establecemos entre la última persona con puntaje 5 y la primera persona con puntaje 4. No existe ningún error.

Finalmente, en el último punto de corte que corresponde al puntaje 0 tenemos 1 error. Así pues, el *item* 1 tiene un total de 6 errores. Dado que hay 30 jueces y 3 categorías, el porcentaje de error es de 6.7 %. Cuando el error es mayor que el 10 % y el *item* permite reagrupación es necesario realizarla, o de lo contrario conviene eliminar al *item*. Se opera de la misma manera con los otros *items*.

Para el cálculo del coeficiente de reproductividad por la Técnica de Cornell, hay que seguir entonces los siguientes pasos: *a)* Determinar los *cutting points* de manera tal de minimizar los errores. *b)* Ningún *cutting point* debe estar tan por encima o tan por debajo que la categoría menor incluya menos del 10 % de los jueces. *c)* Ningún *cutting point* debe ser realizado de manera tal que el número de errores sea mayor que el número de jueces en la categoría.

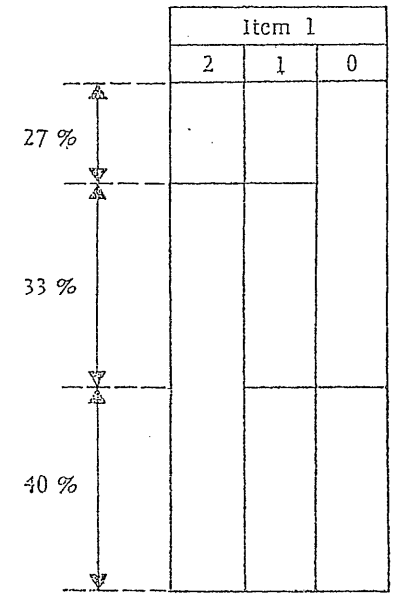
Para controlar si el lector entendió la manera de calcular el coeficiente de reproductividad por la técnica de Cornell, lo invitamos a hacer el cálculo para el Cuadro 10 por su cuenta. Para los *cutting points* indicados en el cuadro, el coeficiente debe ser igual a .949. De hecho, habría que aclarar que el *item* 5 debería ser eliminado, con lo cual el coeficiente de reproductividad aumentaría a un valor aceptable, tal como vimos en el Cuadro 11.

La técnica de Goodenough, es un poco más complicada y se basa en el cálculo de errores en base a pautas marginales.

Los *cutting points* en esta técnica se determinan según la distribución de las respuestas de los jueces a cada una de las alternativas en los distintos *items*. Los cálculos para el Cuadro 10 nos darían los siguientes valores:

	Items												
		1		*2*		*3*		*4*		*5*		*6*	
Categorías	2	1	0	2	0	2	0	2	0	2	0	2	0
Frecuencias	8	10	12	14	16	15	15	14	16	16	14	24	6
Porcentajes	27	33	40	47	53	50	50	47	53	47	53	80	20

Los *cutting points* para cada uno de los *items* deben seguir entonces los porcentajes respectivos. En el caso del *item* 1, los cortes serían entonces:



Es decir, que en el *item* 1, el primer corte o *cutting point* caería entre el primer sujeto con puntaje total 10 y el segundo con puntaje 10, y así sucesivamente para el resto de los *items*.

Mediante estos *cutting points*, se van a determinar pautas de respuestas correspondientes a cada corte. La pauta de respuesta es la manera "correcta" en que deberían distribuirse para cada puntaje total de cada juez si la escala tuviera perfecta escalabilidad. Cada respuesta que no sigue la pauta de respuesta ideal se considera un "error". En nuestro ejemplo, un juez con puntaje de 8 debe seguir una pauta: 0–0–2–2–2–2. Un sujeto con puntaje 6, debe consecuentemente tener una pauta de respuesta 0–0–0–2–2–2. Es decir que primero hay que ordenar a los *items* en términos de escalabilidad, para luego determinar las pautas correspondientes a cada valor. El principio de la técnica es bastante simple. Para 4 *items* dicotomizados y con valores de 1 y 0 para alternativas de respuesta "de acuerdo" y "en desacuerdo", respectivamente, un puntaje total de 3 puede seguir cuatro tipos de pautas diferentes:

a) 0 1 1 1
b) 1 0 1 1
c) 1 1 0 1
d) 1 1 1 0

De estas pautas, solamente la primera es correcta, para *items* que están ordenados en forma acumulativa. Cada una de las pautas *b)*, *c)* y *d)* tiene respectivamente 2 errores (uno por tener un 1 donde debería haber un 0 y otro por tener un cero donde debería tener un 1).

Items													
1		2		3		4		*Puntaje total*	*Pauta de respuesta*				*Errores*
1	*0*	*1*	*0*	*1*	*0*	*1*	*0*						
X		X		X		X		4	1	1	1	1	0
	X	X		X		X		3					0
X			X	X		X		3					2
		X		X			X	3	0	1	1	1	2
	X	X		X		X		3					0
	X	X		X		X		3					0
	X	X		X			X	2					2
	X		X	X		X		2	0	0	1	1	0
	X		X	X		X		2					0
	X		X		X	X		1					0
	X		X		X	X		1	0	0	0	1	0
	X		X		X	X		1					0
	X		X		X		X	0					0
	X		X		X		X	0	0	0	0	0	0
	X		X		X		X	0					0
								Total errores					6

Para hacer más claro el método, en la página anterior dimos un ejemplo más simple, en el que presentamos 4 *items* dicotomizados, las pautas de respuesta y el cálculo de los errores.

El coeficiente de reproductividad, se calcula según la fórmula ya conocida de:

$$r_p = 1 - \frac{\text{Cantidad de errores}}{\text{Número total de respuestas}}$$

En el ejemplo arriba citado:

$$r_p = 1 - \frac{6}{60} = .90$$

El coeficiente de reproductividad nos indica la proporción de respuestas a los *items* que pueden ser correctamente reproducidas.

B) *Ejemplos*

Veamos ahora dos ejemplos, a partir de los cuales presentaremos criterios adicionales al coeficiente de reproductividad para la determinación del universo de contenido en la escala Guttman.

Cuadro de la página siguiente: escalograma Guttman para la medición de nivel de vida determinada en función de materiales empleados en la construcción de viviendas y de las instalaciones sanitarias. Muestra de población de Colombres, en el departamento de Santa Cruz en Tucumán, Argentina. Respuestas de 33 jefes de familia (Arq. Hernández, CES; Universidad de Tucumán, Argentina, 1966).

Referencias

Ejemplo B	*Ejemplo A*
Item 1: Alejamiento de aguas servidas a pozo ciego y fosa séptica.	*Item 1:* Pileta de lavado.
Item 2: Agua corriente.	*Item 2:* Inodoro.
Item 3: Eliminador de residuos.	*Item 3:* Ducha.
Item 4: Techo de cinc, loza o teja con aislación.	*Item 4:* Revestimiento de cemento o superior en baño o letrina.
Item 5: Piso de cemento o mejor.	
Item 6: Luz eléctrica.	

Ejemplo A: Vivienda							*Ejemplo B:* Instalaciones sanitarias				
Sujetos	*Items*						*Sujetos*	*Items*			
	1	*2*	*3*	*4*	*5*	*6*		*1*	*2*	*3*	*4*
1	X	X	X	X	X	X	3	X	X	X	X
33	X	X	X	X	X	X	15	X	X	X	X
31	X	X	X	X	X	X	17	X	X	X	X
22	X	X	X	X	X	X	7	X	X	X	X
19	X	X	X	X	X	X	11	X	X	X	X
16	X	X	X	X	X	X	30	X	X	X	X
4	X	X	X	X	X	X	2	X	X	X	X
6	X	X	X		X	X	1	X	X	X	X
3		X	X	X	X	X	4	X	X	X	X
15		X	X	X	X	X	5	X	X	X	X
17		X	X	X	X	X	8		X	X	X
14		X	X	X	X	X	6		X	X	X
7		X	X	X	X	X	33		X	X	X
2		X	X	X	X	X	31		X	X	X
18		X	X	X	X	X	22		X	X	X
5		X	X	X		X	19		X	X	X
8		X	X		X	X	16		X	X	X
11		X	X	X		X	14		X	X	X
30		X	X	X		X	9				X
29			X		X	X	18				X
9				X	X	X	13				
13				X	X		25				
10				X	X		24				
26					X	X	29				
28					X	X	27				
24					X		26				
27						X	32				
20						X	12				
21						X	20				
25							28				
12							23				
23							10				
32							21				

Cálculo Universo de Contenido para el Cuadro 11:

	Ejemplo A	*Ejemplo B*
Coeficiente de reproductividad:	.949	1.00
Rango marginal mínimo:	.72	.60
Alcance de distribución marginal:	.229	.40

El coeficiente de reproductividad en el Ejemplo B, cuyo valor es 1.00, nos permite decir, sin tener en cuenta la distribución gráfica, que el sujeto 18 tiene en su casa únicamente revestimiento de cemento (o mejor, en su baño), mientras que el sujeto 33, cuyo puntaje es 3, tiene revestimiento, inodoro y ducha, pero no pileta de lavado.

Es muy difícil lograr escalabilidad perfecta, y consecuentemente existen errores que van a ser interpretados como errores de reproductividad. Guttman aconseja que los coeficientes de reproductividad no sean menores de .90.

El coeficiente de reproductividad (r_p) es un criterio necesario, pero no suficiente para la determinación de la escalabilidad de los *items*. Deben tomarse en cuenta otros factores. Stouffer *et al.*,[9] señalan cuatro criterios adicionales: *1)* Alcance de la distribución marginal; 2) Pauta de errores; 3) Número de *items* en la escala, y *4)* Número de categorías de respuestas.

1) *Alcance de la distribución marginal*

Es el más importante de los criterios adicionales, y debe acompañar al coeficiente de reproductividad. El criterio de distribución marginal es determinado por el Rango Marginal Mínimo (M. M. R.) que consiste en el r_p *menos* el promedio de los modos de las frecuencias relativas de las distribuciones de los *items*: $(r_p - MMR)$.

Para algunos, los valores de este criterio adicional deben variar entre .15 y .35; para otros el mínimo debe ser mayor que .10. Estos valores indican *la escalabilidad de los items,* dato que no es proporcionado por el r_p de manera completa (es decir, es posible alcanzar valores altos de r_p —digamos .90— y resultar una escalabilidad inaceptable. Éste es el caso en el cual los *cuttings points* están muy próximos entre sí, con el resultado de discriminar solamente en los extremos de la escala y no a lo largo de la misma. En nuestro ejemplo, los valores de r_p son altos y muy aceptables; los alcances de la distribución marginal, en cambio, son aceptables para el ejemplo *A*, y demasiado altos para el ejemplo *B*.

2) *Pauta de errores*

Cuando el r_p es menor que .90, pero es escalable, es decir que tiene un r_p M. M. R. mayor que .10 estamos en presencia de más de una variable; mejor dicho, de una variable dominante y de otra u otras menores, en el área a través de la cual se ordenan los sujetos. Este tipo de escalograma es denominado cuasi-escala. Éste no es el caso de los dos ejemplos que presentamos.

3) *Número de* items *en la escala*

A mayor número de *items*, mayor la seguridad de que el universo (del cual estos *items* son una muestra), es escalable. Es por esto que cuando los *items* están dicotomizados, como es el caso en nuestros ejemplos, es aconsejable que

[9] Stouffer, S. *et al.: Measurement and Prediction.* Studies in Social Psychology in World War II, vol. IV, J. Wiley & Sons, N. Y., 1966.

su número sea mayor que 10. Pero puede usarse un número menor de *items* si las frecuencias marginales se colocan en un rango con recorridos del 30 % al 70 %.

En los ejemplos dados por nosotros el rango de frecuencias es:

Ejemplo A		*Ejemplo B*	
Item 1	24 %	*Item* 1	30 %
Item 2	57 %	*Item* 2	55 %
Item 3	60 %	*Item* 3	55 %
Item 4	69 %	*Item* 4	60 %
Item 5	78 %		
Item 6	87 %		

De acuerdo al requisito citado más arriba tenemos así alguna seguridad, de que el universo se comporta como la muestra.

4) *Número de categorías de respuestas*

Es otro criterio para asegurar la escalabilidad; cuanto mayor el número de categorías, mayor la seguridad de que el universo es escalable. Por ello, a pesar de la necesidad de reducir las categorías por razones prácticas (disminución del número de errores), hay que asegurarse de que tal reducción no es la resultante de obtener frecuencias marginales extremas (.90–.10) que, como vimos más arriba, no permiten errores, pero artificialmente.

Si mantenemos el número de alternativas de respuestas, a pesar de que aumentará el número de errores, disminuimos la posibilidad de que aparezca una pauta escalable cuando de hecho el universo no lo es.

C) *La técnica de la escala discriminatoria de Edwards y Kilpatrick*

La selección de los *items* iniciales para una escala Guttman ha sido siempre un misterio. En su primer libro Guttman afirma que la selección es cuestión de intuición. Pero, ¿cómo opera esta intuición? ¿No existen reglas o hechos que puedan ser usados cuando se seleccionan los *items* escalables? Aparentemente la intuición de Guttman opera satisfactoriamente porque muchas escalas que han usado esta "técnica" terminan por ser perfectamente escalables. Y aun en las publicaciones más tardías la vía de la intuición parece ser la más correcta.

Si nosotros recapitulamos el proceso total para el investigador, vemos que él, inicialmente, tiene un número de afirmaciones que más o menos miden adecuadamente la variable. De éstos seleccionará un pequeño número al cual aplicará el análisis de escalograma o tal vez la técnica H. Nada indica en la técnica de Guttman que los *items* deben ser elegidos de manera que representen diferentes pasos en el continuo psicológico. Como una cuestión de hecho, todos podrían tener una posición, digamos de 7 en un continuo tipo Thurstone. El único criterio es que los *items* sean escalables de acuerdo con los efectos discriminativos entre un grupo alto y un grupo bajo.

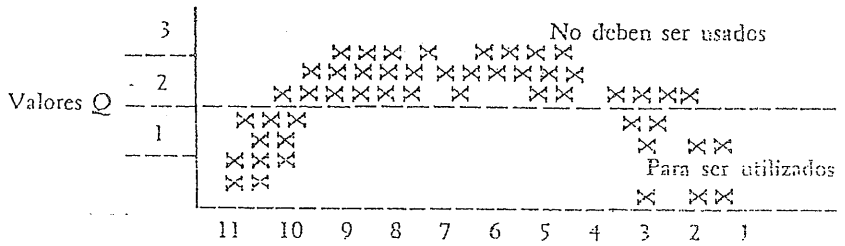
Para superar esto, se distribuyen representativamente los *items* a lo largo del continuo psicológico, y para establecer algunas reglas para la selección de los *items* iniciales de una escala Guttman, Edwards y Kilpatrick desarrollaron una técnica llamada: técnica de la escala discriminatoria *(the scale discrimination technique)*.

Los principales pasos en el procedimiento podrían ser expuestos de la siguiente manera:

1) Por medio de la técnica de Thurstone de los intervalos aparentemente iguales *(the Thurstone equal appearing interval-technique)* un gran número de *items* son colocados a lo largo del continuo.

2) Se calcula para cada *item* la posición mediana y los valores Q.

3) Para disminuir el número de *items* se seleccionan aquellos que en los diferentes intervalos de la escala tienen valores bajos de Q; para esto deben colocar a todos los *items* en una figura como la siguiente:



Se coloca el *cutting point* como lo indica la línea ——— (mediana de la escala Q) y se incluyen aquellos *items* que tienen valores bajos de Q; los que tienen valores altos se excluyen.

4) Ahora se vuelcan los *items* que quedan a una escala Likert, adicionando 5 o 6 alternativas de respuesta para cada uno.

5) Se aplica la nueva escala a por lo menos 100 jueces, y se separan el grupo alto (Q) y el grupo bajo (Q_1), de acuerdo con los puntajes totales.

6) Para probar el poder discriminativo de los *items* se dicotomizan los pesos como en el siguiente ejemplo:

	Grupo bajo	Grupo alto
(5)	2	4
(4)	8 *a*	3 *d*
(3)	7	1
(2)	28	32
(1)	32 *c*	38 *b*
(0)	19	17

La idea es seleccionar el *cutting point* en el continuo de los pesos de manera de minimizar la suma entre las celdas *a* y *d*. Si colocamos el *cutting point* entre (3) y (4) nosotros tendríamos el siguiente resultado:

6	7
86	25

en donde la suma mínima sería $a - d = 31$. Si elegimos el *cutting point* entre (2) y (1) la suma mínima sería igual a 49 y, finalmente, si lo elegimos entre (3) y (2) nosotros encontraremos la suma mínima que es 25. Entonces seleccionamos esta alternativa como la mejor.

7) Después que la selección de las combinaciones mínimas ha sido hecha para todos los *items*, se aplica un test r_ϕ a todos los *items* para seleccionar aquellos que tengan un alto poder discriminatorio. La fórmula a utilizar es la siguiente:

$$r_\phi = \frac{bc - ad}{\sqrt{(a+b)\,(b+d)\,(a+c)\,(c+d)}}$$

donde *a*, *b*, *c* y *d* se refieren a las siguientes celdas:

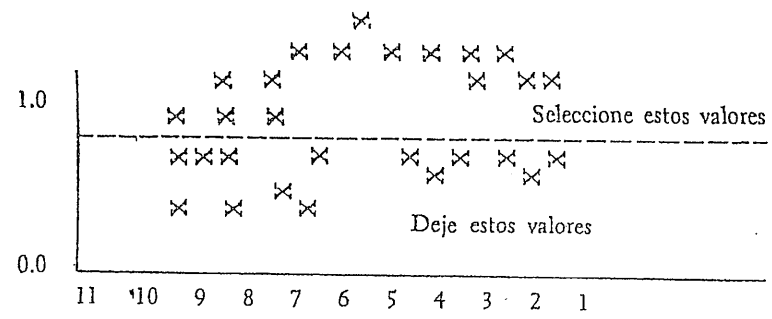
Grupo bajo	Grupo alto	Total
a	*b*	$a + b$
c	*d*	$c + b$
a + c	b + d	

Las figuras que aparecen más arriba reciben el siguiente valor r_ϕ

$$r_\phi = \frac{(87.79) - (17.8)}{\sqrt{(104)\,(95)\,(96)\,(87)}} = .73$$

que es un valor aceptable.

8) Ahora que se tienen los valores r_ϕ para todos los *items* se coloca a todos los *items* en un diagrama para elegir los que tengan valores más altos de r_ϕ. El procedimiento es el mismo que se indicó para el punto 3.



Debe notarse que todo el tiempo tratamos de mantener el continuo psicológico subyacente intocable (manteniendo la distribución de los *items* sobre la escala de 1 a 11 más o menos homogénea).

CUADRO 12

	ITEMS										
Indiv.	*1*	2	3	*4*	*5*	6	7	8	9	*10*	*Score*
1	1	1	4	1	1	3	3	3	4	4	25
2	3	4	3	3	3	4	4	4	4	4	36
3	4	4	4	1	2	2	2	4	4	4	31
4	2	3	1	2	3	4	4	3	3	4	29
5	1	3	2	3	2	4	4	4	4	4	31
6	3	3	3	3	2	4	4	4	4	4	34
7	3	4	3	1	3	3	4	3	4	4	32
8	1	3	1	3	2	4	4	4	4	4	30
9	3	4	3	1	2	4	3	4	4	4	32
10	4	4	4	4	4	4	4	4	4	4	40
11	0	0	0	1	0	3	3	3	0	4	14
12	4	4	4	3	3	4	4	3	4	4	37
13	1	3	1	1	0	2	0	4	4	4	20
14	4	4	4	4	4	4	4	4	4	4	40
15	3	1	1	0	0	2	1	0	3	2	13
16	1	1	1	1	0	2	2	3	2	0	13
17	0	2	2	1	1	3	3	4	4	4	28
18	3	4	4	2	4	4	3	4	4	4	36
19	3	3	4	0	0	4	4	4	4	4	30
20	2	4	3	3	3	4	4	4	4	4	35
21	2	4	3	3	3	3	4	4	4	4	34
22	3	3	4	3	3	4	4	4	4	4	35
23	3	3	4	3	3	4	3	2	2	2	29
24	1	4	4	3	1	4	3	4	4	4	32
25	3	4	3	2	4	4	4	4	0	4	32

D) *La técnica H para mejorar la escalabilidad de la escala Guttman*

Para mejorar la escalabilidad de la escala Guttman se puede utilizar la técnica de H. La idea es formar por intermedio de esta técnica nuevos *items (contrived items)* agrupados con algunos de los *items* iniciales.

Supongamos que tenemos los datos que aparecen en el cuadro 12 (p. 211). Es decir, 25 individuos han respondido a 10 *items*. Nosotros queremos aumentar la escalabilidad formando *"contrived items"*.

1) Existen cinco categorías de respuestas para cada *item* arbitrariamente numeradas 0, 1, 2, 3, 4.

2) Para cada persona se computa un puntaje total basado en todos los *items* que serán sometidos al análisis de escala.

3) Se obtienen las tablas de correlación de cada uno de los *items* considerados con el puntaje total provisorio. Para cada *item* se hacen distintas participaciones, considerando como categorías positivas las respuestas números 1, 2, 3 y 4, después la 2, 3 y 4, después la 3 y 4 y luego la 4.

Empezamos por ejemplo con *item* 1, tomando 3 y 4 como positivo:

CUADRO 13

Puntaje total clasificado	0.1.2	Positivos 3.4
35 – 40	1	6
30 – 34	4	6
25 – 29	3	1
20 – 24	1	
15 – 19		
10 – 14	2	1
5 – 9		
0 – 4		

4) Seleccionamos los *cutting points* para cada *item* que se correlacionan con el puntaje *(score)* total, los cuales son lo suficientemente altos para formar una tabla de 2 por 2 en la cual ninguna celda error tiene una frecuencia mayor que la menor frecuencia de las dos celdas de la diagonal principal Es decir, en este caso:

CUADRO 14

	0, 1, 2	llegamos a: 3, 4
35 – 40 30 – 34	a) 5	b) 12 → Frecuencia Positiva = 12
25 – 29 20 – 24 15 – 19 10 – 14 5 – 9 0 – 4	c) 6	d) 2

Calculamos el r_Q:

$$r_Q = \frac{bc - ad}{\sqrt{(a\,b)\,(b\,d)\,(a\,c)\,(c\,d)}} = \frac{12 \cdot 6 - 5.2}{\sqrt{17 \cdot 14 \cdot 11 \cdot 8}} = .43$$

Se siguen haciendo los mismos cálculos de r_Q para el *item* 1, pero ahora agrupando las categorías de otra manera. Las otras posibilidades son:

	Positivas	
0	1, 2, 3, 4	
0, 1	2, 3, 4	
0, 1, 2	3, 4	(menos utilizado)
0, 1, 2, 3	4	

5) Ordenamos todos los *cutting points* de la frecuencia positiva más alta a la más baja. Véase Cuadro 15.

6) Seleccionamos conjuntos triples de *items* para constituir los cuatro nuevos *contrived items*. El objetivo principal es seleccionar *items* aceptables con la misma frecuencia aproximada para cada triple y espaciar cuanto sea posible estos conjuntos de triples tan extendidos (Cuadro 15).

CUADRO 15.

Item	*Categorías Positivas*	*Frecuencia*	Q		
6	1234	25	.04		
6	234	25	.04		
7	1234	24	.08		
8	1234	24	.08		
8	234	24	.08		
10	1234	24	.08		
10	234	24	.08		
2	1234	24	.08		
3	1234	24	.08		
1	1234	23	.44		
4	1234	23	.44		
7	234	23	.44		
8	34	23	.44		
9	1234	23	.04		
9	234	23	.04		
10	34	22	.62		
10	4	22	.62	*	*Contrived Item I*
2	234	21	.73	*	
6	34	21	.20		
7	34	21	.73	*	
9	34	21	.60		
2	34	20	.80		
3	234	20	.85		
5	1234	20	.80		
9	4	19	.79		
3	34	17	.92	*	*Contrived Item II*
5	234	17	.88	*	
1	234	16	.89	*	
6	4	16	.77		
8	4	16	.53		
4	234	15	.81		
1	34	12	.43		
7	4	13	.63		
2	4	12	.97	*	*Contrived Item III*
4	34	12	.83	*	
5	34	12	.94	*	
3	4	10	.26		
5	4	5	.81	*	*Contrived Item IV*
1	4	3	.53	*	
4	4	2	.37	*	

Reglas: *a)* Si es posible, no repetir los *items* iniciales en diferentes *contrived items*. *b)* Buscar conjuntos de 3 *items* con un máximo de correlación (r_Q).

7) Adjudicar puntajes a cada individuo en cada *contrived item* asignándole un puntaje positivo si él fue positivo en dos o tres de los *items* componentes del *contrived item*.

CUADRO 16. *Contrived items*

Individuos	*I*	*II*	*III*	*IV*
1	−			
2	−	−	−	+
3		+	+	+
4	−	−	+	+
5	−	−	−	+
6	−	−	+	+
7	−	+	+	+
8	−	−	−	+
9	−	−	+	+
10	+	+	+	+
11	−	−	−	+
12	−	+	+	+
13	−	−	−	+
14	+	+	+	+
15	−	−	−	−
16	−	−	−	−
17	−	−	−	+
18	−	+	+	+
19	−	−	+	+
20	−	+	+	+
21	−	+	+	+
22	−	+	+	+
23	−	+	+	+
24	−	+	−	+
25	−	+	+	+

* No escalable

$CR = .96$

$MMR = .76$

Entonces hemos llegado a 4 nuevos *items (contrived items)* que muestran un mayor coeficiente de reproductividad y una mayor diferencia entre CR y MMR.

E) *La versión final de la escala*

Hemos analizado el conjunto total de *items*, llegando finalmente a 10 *items* escalables según los criterios de la escala Guttman. Reproducimos aquí 4 de ellos y la manera de presentarlos en un cuestionario (en el ejemplo que citamos los 10 *items* han sido escalables para 5 valores en las alternativas de respuestas).

1. El patriotismo y la lealtad son los requisitos primeros y más importantes que debe llenar todo buen ciudadano. Ud. está:	( ) Totalmente de acuerdo. ( ) De acuerdo en general. ( ) Ni de acuerdo ni en desacuerdo. ( ) En desacuerdo en general. ( ) Totalmente en desacuerdo.
2. Quien no quiere pelear por su país merece algo mucho peor que la cárcel o los trabajos forzados. Ud. está:	( ) Totalmente de acuerdo. ( ) De acuerdo en general. ( ) Ni de acuerdo ni en desacuerdo. ( ) En desacuerdo en general. ( ) Totalmente en desacuerdo.
3. Cualquier esfuerzo hecho en el entrenamiento militar de Chile, es compensado por los beneficios de su seguridad. Ud. está:	( ) Totalmente de acuerdo. ( ) De acuerdo en general. ( ) Ni de acuerdo ni en desacuerdo. ( ) En desacuerdo en general. ( ) Totalmente en desacuerdo.
4. Para Chile sería un grave error permitir el ingreso de los extranjeros que quitan las oportunidades de trabajo a los nacionales. Ud. está:	( ) Totalmente de acuerdo. ( ) De acuerdo en general. ( ) Ni de acuerdo ni en desacuerdo. ( ) En desacuerdo en general. ( ) Totalmente en desacuerdo.

La escala es unidimensional, es decir, que los resultados obtenidos de los sujetos van a permitir ubicarlos en el continuo favorable-desfavorable a la actitud, en un rango que va (en nuestro ejemplo) de 0 a 40 puntos.

F) *Ventajas y desventajas de la escala Guttman*

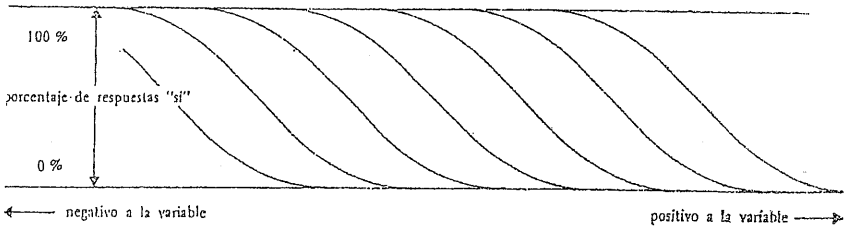
Ventajas. a) Se asegura en forma casi definitiva la unidimensionabilidad de la escala. *b)* Conociendo el puntaje de un individuo se puede saber qué grado de acuerdo tuvo con los *items* y ubicarlo así en el continuo de la escala.

Desventajas. a) Cuando se trata de medir actitudes complejas conviene hacer un tipo de escala para cada dimensión de la actitud. *b)* Involucra mayor cantidad de trabajo que en las dos escalas mencionadas anteriormente (Lickert y Thurstone). *c)* Escalas de este tipo pueden resultar unidimensionales para un grupo y no para otros grupos.

G) *Comentarios finales*

Los *items* de la escala Guttman, como los de la escala Lickert, son acumulativos. Sin embargo, los *items* en la escala Lickert son, cada uno de ellos, acumulativos; mientras que en la escala Guttman consiste en una *serie acumulativa* de *items*, donde cada *item* tiene además carácter de acumulativo. De allí que en la escala Lickert sea posible que todos los *items* puedan ocupar aproximadamente el mismo lugar de la escala; en el caso de la escala Guttman, utilizando la técnica H de Edwards y Kilpatrick, garantizamos que los *items* en la escala se distribuyan a lo largo de el continuo de actitud. En términos gráficos:

Figura 8.



Ejemplo de algunos *items* acumulativos que forman una serie acumulativa de *items*. Todos los *items* van en la misma dirección (todos los *items* son de tal tipo que todos los sujetos con una disposición negativa respondan "Sí"). En realidad la inclinación para cada *item* y las distancias entre los *items* varían.

¿Variación en el instrumento, en los sujetos, o en ambos?

La escala Guttman ordena tanto a los sujetos como a los estímulos con respecto a un continuo de actitud, es decir que tanto a sujetos como a estímulos se les puede asignar valores de escala. Consecuentemente, la escala Guttman corresponde a un enfoque centrado en la respuesta. En este caso, tanto la actitud del sujeto como la actitud reflejada por el estímulo, actúan para determinar la respuesta del individuo.

MÉTODO DE COMPARACIÓN POR PARES

Consiste en presentar los estímulos al sujeto de a dos por vez (pares) y preguntarle cuál de ellos es el más grande, el mejor, el más cálido, etc. Por medio de este artificio logramos un orden de rango basado en el número de elecciones recibidas por cada *item*.

Como las comparaciones son sistemáticas (el sujeto compara todo estímulo con cada uno de los otros), el resultado final nos indicará cuán consecuente es el individuo con sus juicios o evaluaciones. En otras palabras, esto nos da la consistencia de las respuestas o la confiabilidad. En una pauta totalmente consistente, debemos esperar que el estímulo que sea más grande, más agradable, más cálido, etc., para el sujeto tendrá $n-1$ elecciones, el segundo estímulo $n-2$ elecciones, y así sucesivamente hasta llegar al menos grande, que tendrá 0 elecciones.

Se debe tener cuidado con no presentar sistemáticamente un estímulo en primer o segundo lugar, ya que con ello podríamos favorecer o contrariar la elección. Para evitar esto, se colocan los estímulos aleatoriamente, o se administran formas paralelas.

Este método de comparación por pares requiere de mucho tiempo cuando las comparaciones a realizar son numerosas. El número posible de pares a calcular resulta de la fórmula

$$\frac{n\,(n-1)}{2}$$

donde n = cantidad de *items*.

Es decir que si queremos comparar 50 *items*, el número de combinaciones que el sujeto deberá evaluar son: 1 225; si utilizamos 100 *items*, el número de pares a comparar serán 4 950.

A) *Ejemplo de construcción de una escala basada en el método de comparación por pares*

El método de comparación por pares puede ser usado como un *test* de Tipo A, o como un *test* de Tipo B. Analizaremos aquí los resultados de un *test* de Tipo B en el caso de los jueces y como un *test* de Tipo A en la aplicación a un sujeto.

1) *Versión de los jueces.* Test *de Tipo B*

Supóngase que queremos determinar el grado relativo de izquierdismo-derechismo de 5 partidos políticos en un país cualquiera, para después poder ordenar el grado de izquierdismo-derechismo en sujetos según sus simpatías políticas por partidos.

En la versión de los jueces vamos a ubicar a los partidos en un orden de rangos, y además —según la evaluación de los jueces— en una escala intervalar en donde un extremo de la escala representará izquierdismo y el otro derechismo.

A un grupo de 50 jueces solicitamos que ordenen los 5 partidos *A, B, C, D* y *E* dando un valor 0 al partido más a la izquierda, un valor 1 al siguiente, etc., hasta un valor 4 al partido ubicado más a la derecha del espectro político.

Analizaremos los resultados de 5 jueces (se entiende que la evaluación de la escala debe hacerse sobre el total de jueces que se utilicen), para simplificar las operaciones. Los resultados de los ordenamientos de nuestros 5 jueces son los siguientes:

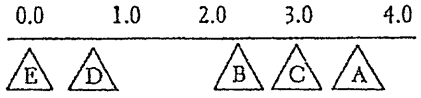
Juez	*Ordenación de los partidos*				
	A	C	B	D	E
1	4	3	2	1	0
2	3	2	4	0	1
3	2	4	3	1	0
4	4	2	3	1	0
5	3	4	2	1	0

El puntaje máximo posible con 5 jueces para un partido es $\frac{20}{5}=4$. El puntaje mínimo es 0.

Los puntajes para los cinco partidos son entonces:

$$A = (4+3+2+4+3)/5 = 3.2$$
$$C = 15/5 = 3.0$$
$$B = 14/5 = 2.8$$
$$D = 4/5 = 0.8$$
$$E = 1/5 = 0.2$$

El orden de rango de los partidos, de derecha a izquierda es: $A-C-B-D-E$, y su ubicación en una escala intervalar:



2) *Versión de la escala para sujetos. Versión final.* Test *de Tipo A*

En el cuestionario se presentan los partidos políticos por pares, en una lista

en la que se combinen de dos en dos todos los partidos. En nuestro caso los pares a comparar serían:

A-B	B-D
A-C	B-E
A-D	C-D
A-E	C-E
B-C	D-E

Cuidamos por supuesto que los partidos aparezcan "mezclados" en la lista, evitando por ejemplo que el partido A esté siempre al comienzo, y se le solicita al sujeto seleccione entre cada par a cuál de los dos partidos prefiere.

B) *Ejemplo de una escala de comparaciones por pares en un cuestionario (Entrevista)*

El entrevistador muestra una tarjetita (Hoja III) separada con el contenido indicado con la línea —.———— arriba. En el cuestionario aparece el texto total arriba. El entrevistador sigue las instrucciones preguntando cada vez por un solo par.

Las indicaciones se hacen naturalmente en el espacio en el cuestionario y no en la tarjeta (Hoja III). (El entrevistador hace la registración en el cuestionario señalando el entrevistado solamente su preferencia.)

Un entrevistado ha contestado de la siguiente manera:

35. Tengo aquí una hoja con los nombres de los partidos políticos, agrupados de dos en dos. Por favor, dígame cuál partido prefiere en cada grupo.

Muestre hoja III

Subraye el partido escogido, si dos partidos son escogidos al mismo tiempo, subraye a ambos Si el entrevistado no puede escoger ningún partido, no subraye nada.

Conservador – Liberal.
Socialista – Comunista.
Liberal – Partido Laborista.
Laborista – Socialista.
Laborista – Conservador.
Comunista – Laborista.
Conservador – Socialista.
Liberal – Comunista.
Socialista – Liberal.
Comunista – Conservador.

Conserv. 1	= 1
Lib. 2 + 2 + 2 + 2	= 8
Cent. 2 + 2 + 1	= 5
Soc. 2 + 2	= 4
Com. 2	= 2

Las ponderaciones son calculadas después por el entrevistador. Se pone 2 si una de las alternativas fue tomada, y 1 si ambas fueron tomadas. El resultado es el perfil 1, 8, 6, 4, 1 del sujeto. Podemos concluir que la persona simpatiza más con el partido liberal. La ponderación 2 y 1 que corresponden a cada par es arbitraria. Se puede utilizar, por ejemplo, también 1 y 0. En este caso el perfil sería 0, 4, 3, 2, 0, y el sujeto todavía simpatiza más con los liberales.

Aparte de ser posible determinar la posición del sujeto en el *continuum* izquierda-derecha, se puede, por ejemplo, estudiar las personas que tienen *inconsistencia* en sus perfiles. Además del perfil obtenido podemos calcular la posición del individuo según nuestra escala:

Resultados para el individuo

Posición del partido a lo largo de la variable	*Nombre del partido*	*Columna M Núm. de elecciones "recibidas"*	*Posición del individuo a lo largo de la variable*
++ 4	A	1	4 por 1 = 4
+ 3	C	8	3 por 8 = 24
2	B	5	2 por 5 = 10
— 1	D	4	1 por 4 = 4
— 0	E	2	0 por 2 = 0
			Total = 42

El individuo tiene una posición medianamente en el centro de la escala. Una persona extrema en la variable recibe como máximo 60 puntos y como mínimo 20. Compruebe esto usted mismo a través de cálculos.

Es evidente que es mucho más fácil para el sujeto indicar cuál de las 2 alternativas prefiere más. Sopesar cada *item* en una escala es más difícil como también es complicada la tarea de ordenar por rango los *items*. Cuando el número de *items* es más de 3 ó 4 en la tarea de ordenar por rango, el sujeto mismo tiene que usar alguna especie de comparación de los *items* entre sí. De este modo el método de comparaciones por pares se hace en una forma sistemática. El resultado de este método es un orden de rango basado en el número de elecciones recibidas por cada *item*.

Una ventaja con el método de comparaciones por pares es que muestra cuán "consecuente" es el sujeto en su estimación de las alternativas comparadas. Una inconsecuencia se muestra al disminuir irregularmente la serie de números en la columna *M*. Si el número de *items* es *N* y si el sujeto es consecuente, nosotros obtendremos $(N-1)$ elecciones recibidas para el *item* que al sujeto le gusta más, $(N-2)$ para su segunda preferencia de *item*, etc. Si el número de *items* es 7, así recibimos 6, 5, 4, 3, 2, 1, 0 (en el caso de la persona mencionada anteriormente). 6, 5, 4, 2, 2, 2, es un ejemplo de inconsecuencia. La suma de los valores de la serie de números es en ambos casos 21, el número de comparaciones por pares.

Hay algunas variantes del método de comparaciones por pares, como el método de comparaciones por pares dobles y comparaciones tríadas que han sido usadas en medición de intereses y actitudes. Estos métodos sin embargo, hacen más difícil que el sujeto haga una ordenación "consecuente" (precisión más baja).

Algunos *test* de interés y actitudes muy conocidos que usan el método de orden por rango o el método de comparaciones por pares son: el *test* de interés de Allport Vernon, Lindzey en *"Study of Values"* (actitudes hacia la religión, política, arte, ciencia, etc.) y el registro de preferencia Kuder (interés en actividades como deporte, música, etc.).

C) *Ventajas y desventajas*

Ventajas. a) Proporciona resultados más precisos. *b)* Muestra claramente cuán "consecuente" es un sujeto con sus propios juicios. *c)* Es más fácil para un sujeto indicar a cuál de dos alternativas prefiere más.

Desventajas. a) Cuando los juicios son numerosos, las comparaciones son muy grandes. En el caso de 20 *items*, el número de comparaciones a realizar es de 190. *b)* Gran consumo de tiempo de entrevista.

El diferencial semántico. La escala de Osgood

A) *Antecedentes teóricos*

El método es descrito por los autores como un método para medir el significado que tiene un objeto para un individuo.

Osgood supone que existe un espacio semántico de dimensionalidad desconocida y de naturaleza geométrica. El espacio está construido (o constituido) de escalas semánticas. Cada escala consiste de un par de adjetivos que son bipolares. Se supone que estas escalas forman una función lineal que pasa a través del origen. Para estar en condiciones de definir el espacio adecuadamente, es necesario usar una gran cantidad de escalas que son una muestra representativa extraída del universo de escalas. Para diferenciar el significado de un objeto, el individuo hace una elección entre las alternativas dadas. La función de cada elección es localizar el objeto en el espacio semántico. La validez de la localización en este punto en el espacio depende del número y representatividad de las escalas.

De este modo, la diferencia semántica significa la estabilización sucesiva (anclaje) de un objeto hasta un punto en el espacio multidimensional semántico, a través del puntaje de un número de alternativas semánticas dadas presentadas en la forma de escalas. Una diferencia de significado entre 2 objetos es simplemente una función de las diferencias de su ubicación en el mismo espacio, es decir, una función de la distancia multidimensional entre 2 puntos.

El punto en el espacio que da una definición operacional del significado tiene 2 características principales: *1)* Dirección desde el origen; 2) Distancia desde el origen. Esto podría ser explicado como el *tipo* e *intensidad* del significado.

La *dirección* desde el origen depende de cuál de los polos de la escala se elige y la *distancia* depende de cuán extrema es la posición elegida en la escala.

B) *Dimensiones en el espacio semántico*

Osgood dio gran importancia al muestreo. El diferencial semántico está influido por 3 fuentes de variación: el individuo, las escalas y los objetos. Muchas diferentes modificaciones fueron hechas para asegurar la universalidad de la estructura del factor *(factor-structure)* pero siempre Osgood obtuvo los mismos factores principales en los diferentes análisis y así llegó a la conclusión de que la estructura del factor no dependía de la elección de escalas. El seguir 3 factores de hecho explicó la mayor parte de la varianza total, mientras otras dimensiones sólo explicaban una pequeña parte de ellas.

Dimensiones

1) La *evaluación* que hace el individuo del objeto o concepto que se está clasificando. Ejemplo de escalas bipolares: regular-irregular; limpio-sucio; bueno-malo; valioso-despreciable.

2) La percepción del individuo de la *potencia* o poder del objeto o concepto. Escalas: grande-chico; fuerte-débil; pesado-liviano.

3) La percepción del individuo de la *actividad* del objeto o concepto. Escalas: activo-pasivo; rápido-lento; frío-caliente.

C) *Construcción*

El método para el diferencial semántico no es una prueba con ciertos *items* y puntajes de *tests* específicos. Debe ser visto como un método para reunir cierto tipo de información (un método que puede ser generalizado), el cual tiene que constituirse por las demandas que presenta cierto problema de investigación. No hay objetos estándar o escalas estándar.

Selección del objeto (concepto): "objeto" se usa para determinar qué significa el "estímulo" que da "reacción" (respuesta) en el individuo a través de su indicación en las escalas de adjetivos.

El objeto puede ser verbal; puede consistir de sólo una palabra o de varias palabras. Objetos no-verbales pueden ser diferenciados (cuadros u otros estímulos estéticos).

La elección correcta de un objeto es un problema de muestreo. Esto generalmente significará en la práctica común que el investigador usa su sentido común al seleccionar el objeto. El investigador debería pensar en elegir objetos que se supone darán: *1)* Diferencias individuales (para poder estudiar la variación en el material). *2)* Que tengan un solo significado (de otra manera hay riesgo de vacilación en la elección). *3)* Se supone que todos los individuos lo conozcan bien (de otro modo habrá regresión al medio de la escala).

D) *Selección de escalas*

Al seleccionar las escalas se debe estar seguro de tener los 3 factores: evaluación, potencia y actividad representadas. Para cada factor uno puede seleccionar cierta cantidad de escalas correlacionadas y a través de éstas obtener el promedio de las escalas. Esto aumentará la representatividad y la confiabilidad. Los promedios son llamados *factor scores*. El criterio más importante para la selección de escalas así es su patrón factorial (generalmente se seleccionan 3 escalas para cada factor. Las escalas elegidas cargadas en ese factor son bajas en los otros).

Otro criterio es la relevancia de las escalas al objeto. Las escalas con composición factorial desconocida pueden ser usadas si se suponen ser muy relevantes para la investigación. En este caso uno puede incluir escalas con composición factorial conocida y usarlas como *escalas de referencia* y estudiar cómo las escalas "desconocidas" covarían con las escalas de referencia

Análisis de los datos

Se les deberán asignar pesos a las diferentes alternativas de la escala: 1, 2, 3, 4, 5, 6, 7; o 7, 6, 5, 4, 3, 2, 1, dependiendo de cuál polo en el par de adjetivos debe ser considerado como alto para este factor.

Por ejemplo:

bueno	malo	o	bueno	malo

dependiendo del orden de presentación. Las escalas dentro del mismo factor se presentan como variable positiva y negativa como en la escala Lickert. Usando los pesos es posible hacer cálculos aritméticos corrientes, perfiles, análisis factorial, etc.

E) *El* test *de Osgood y la medición de actitudes*

Tal como se ha hecho notar, el *test* de Osgood no es una escala de actitudes en el sentido corriente, sino que las investigaciones han demostrado que los valores de escala pueden ser usados para la medición de actitudes. La actitud general se mide pero no el contenido de la actitud en el significado denotativo.

Osgood piensa que con la ayuda de los valores de escala es posible formar una escala universal que podría ser usada para medir la actitud hacia cualquier objeto. La ventaja con tal escala universal sería: 1) Económica (tiempo, dinero, trabajo). 2) Disponible y así hacer posible la medición de actitudes que no habían sido previstas. 3) Posibilidades de comparar entre diferentes investigaciones de actitud y actitud hacia diferentes objetos.

A continuación presentamos un *ejemplo*:

ESCALAS PARA LA MEDICIÓN DE ACTITUDES

INSTRUCCIONES

Les presentamos algunos párrafos tomados de periódicos y revistas que indican una relación entre diferentes asuntos de interés. Su tarea consistirá en evaluar dichos párrafos de acuerdo a una serie de escalas de adjetivos que encontrará a continuación del texto que *usted* debe juzgar.

El procedimiento consiste en señalar con una X en la escala el lugar en el cual usted ubicaría *su* opinión acerca del texto.

Por ejemplo: La palabra *fanático* podría ser evaluada en las siguientes escalas:

flexible	___:	___:	___:	___:	___:	___:	___	rígido
negativo	___:	___:	___:	___:	___:	___:	___	positivo
delicado	___:	___:	___:	___:	___:	___:	___	rudo

Si alguien considerara que la palabra *fanático* indica rigidez, debería marcar en la escala el lugar que está más próximo al adjetivo rígido, de este modo:

flexible	___:	___:	___:	___:	___:	___:	X ___	rígido

Si considerara que indica rigidez pero en menor grado, marcaría en:

flexible	___:	___:	___:	___:	___:	X ___:	___	rígido

o en cualquier otro lugar menos cercano al adjetivo rígido, según lo considerara. Lo mismo haría con cada una de las otras escalas. *Haga usted lo mismo con cada uno de los textos.*

Recuerde que debe señalar *su* opinión. No deje de hacerlo en ninguna de las escalas.

Si tiene preguntas formúlelas antes de comenzar a trabajar. Es conveniente que trabaje en silencio y lo más rápidamente posible.

Porque obviamente se ve, que una América Latina desarmada espiritual y materialmente ante la amenaza creciente y reciente de la Conferencia Tricontinental, quedará atada de pies y manos para la invasión, la subversión y el consiguiente desembarco de infantería de marina en la lucha anticomunista posterior, inmediata e inevitable.

bueno	___:	___:	___:	___:	___:	___:	___	malo
débil	___:	___:	___:	___:	___:	___:	___	fuerte
exaltado	___:	___:	___:	___:	___:	___:	___	moderado
absurdo	___:	___:	___:	___:	___:	___:	___	significativo
blando	___:	___:	___:	___:	___:	___:	___	duro
pasivo	___:	___:	___:	___:	___:	___:	___	activo

positivo	___:	___:	___:	___:	___:	___:	___	negativo
poderoso	___:	___:	___:	___:	___:	___:	___	impotente
frío	___:	___:	___:	___:	___:	___:	___	cálido

La mísera situación social de la minoría negra, que vive hacinada en casas de madera de los barrios extremos, ha rebelado una vez más a la gente de color que constantemente lucha por mejorar sus condiciones de vida.

Frente a los aparatos de refrigeración, el aire acondicionado, la vivienda sana y cómoda de los blancos, los negros sólo disponen de las bocas de riego para refrescarse del espantoso calor de los últimos días. La policía no atendió razones y cerró las bocas de riego, pero los negros enfurecidos atacaron a los intrusos y esto dio origen a una más de las tantas tragedias raciales que se producen en los Estados Unidos.

bueno	___:	___:	___:	___:	___:	___:	___	malo
débil.	___:	___:	___:	___:	___:	___:	___	fuerte
exaltado	___:	___:	___:	___:	___:	___:	___	moderado
absurdo	___:	___:	___:	___:	___:	___:	___	significativo
blando	___:	___:	___:	___:	___:	___:	___	duro
pasivo	___:	___:	___:	___:	___:	___:	___	activo
positivo	___:	___:	___:	___:	___:	___:	___	negativo
poderoso	___:	___:	___:	___:	___:	___:	___	impotente
frío	___:	___:	___:	___:	___:	___:	___	cálido

El mercado bursátil

Poca actividad y precios irregulares mostró durante las operaciones de ayer el mercado bursátil. Los mejores conformes pertenecieron a Banco de Chile, Tierra del Fuego, Copec, Melón, Papeles y Cartones, Cap. y Mademsa.

Los bonos se anotaron en pequeñas partidas de Deuda Interna 7-1 y Banco Hipotecario de Chile 10-6 con cotizaciones mantenidas.

El volumen de los negocios ascendió a la suma de E° 331 nominales en bonos y 279.582 acciones.

El movimiento en transacciones se vio poco activo siendo inferior al de la semana próxima pasada.

A la hora del cierre de las operaciones, la plaza mostraba poca actividad y precios irregulares.

bueno	___:	___:	___:	___:	___:	___:	___	malo
débil	___:	___:	___:	___:	___:	___:	___	fuerte
exaltado	___:	___:	___:	___:	___:	___:	___	moderado
absurdo	___:	___:	___:	___:	___:	___:	___	significativo

blando	___:	___:	___:	___:	___:	___:	___	duro
pasivo	___:	___:	___:	___:	___:	___:	___	activo
positivo	___:	___:	___:	___:	___:	___:	___	negativo
poderoso	___:	___:	___:	___:	___:	___:	___	impotente
frío	___:	___:	___:	___:	___:	___:	___	cálido

Escala de distancia social de Bogardus

Para finalizar nuestro capítulo sobre escalas, presentaremos la escala de distancia social de Bogardus, quien de hecho fue el primer científico social en utilizar el concepto de escala para la medición de actitudes. La escala de distancia social fue utilizada por Bogardus para el estudio de las relaciones —manifestadas en términos de actitudes— entre grupos étnicos y nacionales. Ulteriormente la utilización de este tipo de escala se ha extendido al estudio de relaciones entre clases sociales, grupos ocupacionales, etc.

El concepto de distancia social se refiere al grado de conocimiento o simpatía o intimidad que caracterizarían las relaciones entre distintos grupos. Por medio de ella es posible medir el grado de "aceptabilidad" que un grupo social cualquiera tiene para otro grupo. Los *items* a partir de los cuales se construye la escala están graduados de manera tal que van de la aceptación (medida por un grado alto de aceptación de relaciones íntimas) al rechazo (medida por la negación a cualquier relación).

A) *Procedimientos básicos para la construcción de una escala de distancia social*

1) Defina el continuo —dentro del sentido del concepto de distancia social arriba mencionado— en términos que van desde un contacto estrecho, afectuoso e íntimo, pasando por la indiferencia, hasta llegar a la antipatía, la hostilidad y el rechazo.

2) Reúna una serie de *items* de distancia social igualmente espaciados. Tales *items* constituyen afirmaciones que se refieren a un gran número de relaciones sociales.

3) De acuerdo con el método de Thurstone —unas 100 personas que hacen de jueces— las juzgan en términos de grados crecientes de distancia social o de grados crecientes de proximidad social.

4) En el estudio de Bogardus, siete de estos *items* igualmente espaciados, de grados crecientes de distancia social según la determinación de los jueces, fueron seleccionados.

5) A estos siete *items* se les dan números arbitrarios que van de 1 a 7, respectivamente, representando creciente distancia social en esta proporción.

Debemos aclarar que Bogardus, para la selección de los *items*, muchas veces ha empleado su experiencia; es decir que no ha recurrido a jueces para

que, de acuerdo a las evaluaciones de éstos, sean seleccionados finalmente los *items* considerados adecuados para que integren la escala.

En la práctica, lo que han hecho los investigadores es aplicar directamente la escala ya estructurada por Bogardus, consistente en 7 *items*. O, por el contrario, solamente han adaptado tales *items* —en sus formulaciones— a las necesidades del estudio.

B) *A continuación presentamos los* items *que ilustran esta escala de distancia social*

ESCALA DE DISTANCIA RACIAL

Instrucciones

1) Dé su primera reacción ante cada pregunta sin pensarlo demasiado.

2) Dé sus reacciones a cada raza considerada como un grupo. No dé sus reacciones de sentimientos a los mejores o a los peores miembros que usted ha conocido, sino piense de acuerdo a la idea que usted tiene de la raza considerada como un todo.

3) Ponga una cruz debajo de cada raza en tantas de las siete filas como sus sentimientos lo dicten.

Categoría (*item*)	Judío	Negro	Indio	Blanco europeo	Chino	Mestizo
1. Se casaría con	.....	.....	.....	.....	.....	.....
2. Tendría como amigos regulares	.....	.....	.....	.....	.....	.....
3. Trabajaría en una oficina junto a	.....	.....	.....	.....	.....	.....
4. Viviría en el mismo vecindario con	.....	.....	.....	.....	.....	.....
5. Estaría simplemente hablando como a conocidos a	.....	.....	.....	.....	.....	.....
6. Excluiría de mi vecindario a	.....	.....	.....	.....	.....	.....
7. Excluiría de mi país a	.....	.....	.....	.....	.....	.....

El sujeto es, pues, interrogado para dar sus primeras reacciones de sentimientos y no para racionalizar. La suposición es que las primeras reacciones de sentimientos reflejan actitudes mejor que ningún otro, aparte de la conducta en sí en un periodo.

Aunque la conducta de larga duración es la mejor prueba de las actitudes de una persona, la escala de distancia social se ideó para dar predicciones, mientras se espera que la conducta de larga duración revele las actitudes. La conducta de un corto periodo puede revelar seudoactitudes, no actitudes verdaderas. Ella puede medir actitudes que son "ocultas" para ciertos fines.

Se supone, en consecuencia, que las primeras reacciones de sentimientos sin racionalización son significativas en revelar cómo actuaría una persona si ella tuviera que enfrentarse repentinamente con las situaciones citadas en la escala.

C) *Flexibilidad de la técnica*

Otra indicación de la flexibilidad de esta técnica puede verse en el hecho de que ella puede ser usada no solamente para escalar grupos o valores que son externos al sujeto que hace el *rating*, sino también para escalar a los mismos *raters* con respecto a la distancia social entre ellos y algún grupo.

Para ello se utiliza un *continuum* en el sentido de favorabilidad-desfavorabilidad de la escala tipo Lickert. Con un ejemplo se ilustrará mejor:

INSTRUCCIONES

1) Hablando de los norteamericanos, por favor, ¿podría indicarnos su simpatía o no hacia ellos?

2) Por favor, evalúe a los norteamericanos en esta escala, marcando con una cruz en los espacios punteados la afirmación que expresa su sentimiento hacia ellos.

No les tengo ni simpatía ni antipatía	Me son decididamente simpáticos	Me son un poco simpáticos	Decididamente me son antipáticos	Les tengo un poco de antipatía
.....	.....	.....	.....	.....

Esta forma de escala es recomendable y se presenta primero a los sujetos a quienes se les está sometiendo a investigación acerca de sus actitudes. Luego viene la forma anteriormente diseñada de la escala.

La forma ilustrada anteriormente se construye de la misma manera para los demás grupos étnicos, o nacionales, hacia quienes deseamos conocer las actitudes de los sujetos.

D) *Confiabilidad*

Una escala de distancia social no es fácilmente probada para la determinación de su confiabilidad, ya sea por la forma múltiple o por la técnica de la división por mitades *(split-half)*. El enfoque *test-retest* es la medida más efectiva de confiabilidad de tal escala.

E) *Validez*

Para mostrar la validez de la escala se requiere de pensamiento cuidadoso. La aplicación del método del *known-group* implicaría hallar grupos conocidos que sean favorables hacia algunos de los tipos étnicos y no favorables hacia otros. Si las respuestas de estos grupos forman el requisito patrón, entonces la validez parecería probable.

Por otro lado, el uso del método de criterios independientes requeriría que el orden de rango forme algún otro rango de aceptabilidad social. Tales indicadores podrían ser el orden de rango de deseabilidad como inquilinos en un gran proyecto de construcciones de viviendas, de aceptabilidad como miembros de un gremio, etc.

F) *Limitaciones y aplicaciones*

Esta técnica de *scaling* no está limitada en cuanto a su flexibilidad de aplicación ni por su crudeza como medida. Los principales problemas son: *1)* La suposición de la equidistancia entre los puntos de la escala; *2)* La suposición de que cada punto está necesariamente "más allá" del punto anterior; *3)* El hecho de que ella puede ser probada por confiabilidad solamente por el "decir" de la técnica ordinaria del *test-retest*. Por lo tanto, el uso de este método de *scaling* está limitado generalmente a estudios pilotos o a investigaciones que por alguna razón deben ser completados rápidamente y no requieren de un nivel de precisión muy alto.

Bibliografía recomendada para escalas de medición de actitudes

Al lector interesado en profundizar tanto en la teoría como en la técnica de escalas de medición de actitudes recomendamos especialmente los siguientes textos:

Edwards, A. *Techniques of Attitude Scale Construction;* Appleton-Century-Crofts, Nueva York, 1957.

Torgerson, W. *Theory and Methods of Scaling;* J. Wiley, Nueva York, 1958.

Thurstone, L. *The measurement of Attitudes,* The University of Chicago, 1929.

Stouffer, S. *et al. Measurement and Predication. Studies in Social Psychology in World War II;* Princeton University, 1950.

Upshaw, H. "Attitudes Measurement" en Blalock, H. y Blalock, A. *Methodology in Social Research;* McGraw-Hill, Nueva York, 1968.

Selltiz, Jahoda *et al. Research Methods in Social Relations;* Holt-Rinehart-Winston, Nueva York, 1959.

Recomendamos también revisar especialmente las siguientes revistas:

Sociology and Social Research
Journal of Applied Psychology
Journal of Social Psychology
Journal of Abnormal and Social Psychology
American Sociological Review
American Journal of Sociology
Psychological Bulletin

VII. EL TRABAJO DE CAMPO

Ingvar Ahman

Supongamos que los entrevistadores han sido seleccionados y se ha verificado su capacidad para el trabajo en el terreno en el cual va a ser realizado. Supongamos también que el investigador cuenta con una central para el trabajo de terreno, desde la cual él opera. Esta central tiene servicio telefónico, una gran sala de conferencias y algunas salas pequeñas, a las cuales los entrevistadores pueden llevar a los *R*,[1] si es que existen dificultades para obtener un lugar adecuado para realizar la entrevista en otra parte. En la central existen además planos de la ciudad, guías telefónicas, mapas de comunicaciones, etc.

Algunos días antes de que comience el trabajo en terreno, los entrevistadores deben tener una reunión preliminar, con el objeto de tomar todas las instrucciones necesarias, ejercitarse en algunos cuestionarios, formularios, etc. Algunos de los puntos de importancia que deben tratarse en estas reuniones preliminares, son los siguientes:

1) *Descripción del estudio*

El investigador ofrecerá una descripción metodológica del estudio y explicará algunas de las partes teóricas que hay dentro de él.

2) *Descripción de la muestra*

El investigador dará una explicación acerca de la forma como se ha diseñado la muestra, su tipo y qué posibilidades existen para reemplazar los individuos que no se encuentren a) efectuar las entrevistas.

Los puntos 1 y 2 son meramente informativos para ubicar a los entrevistadores en el tipo de estudio y si el investigador lo considera conveniente, pueden eliminarse. Los puntos que vienen a continuación son bien específicos y conviene se los tenga en cuenta.

3) *Cómo ponerse en contacto con los respondentes*

A cada entrevistador debe asignársele un distrito especial, propio. El entrevistador recibe un mapa y se le pide que haga un plan para sus contactos, que no ofrezca mucha pérdida de tiempo en movilización. De acuerdo con la profesión del *R*, el entrevistador selecciona el tiempo que considera más probable para contactarse con las personas elegidas. Es importante que se contacte con el mayor número de *R* posibles. No siempre existe la seguridad de que la persona disponga del tiempo necesario para la entrevista cuando es visitado por el entrevistador, aunque esté realmente interesado. En este

[1] *R* = Respondente.

Esta obra se terminó de imprimir el día 29 de enero de 1982 en los talleres Bolea de México, S. A., Calle 3, núm. 9-A, Naucalpan de Juárez, Estado de México.

La edición consta de 10 000 ejemplares y sobrantes para reposición.