

ANUARIO DE PSICOLOGÍA  
Núm. 43 - 1989 (4)

# FIABILIDAD Y GENERALIZACIÓN DE LA OBSERVACIÓN CONDUCTUAL

ÁNGEL BLANCO VILLASEÑOR  
Departamento de Metodología  
de las Ciencias del Comportamiento  
Universidad de Barcelona

Ángel Blanco Villaseñor  
Departamento de Metodología de las Ciencias del Comportamiento  
Facultad de Psicología  
Adolf Florensa, s/n.  
08028 Barcelona

## *De la fiabilidad del registro observacional a la generalizabilidad de la observación*

En las ciencias del comportamiento los fenómenos observados están influidos por tal cantidad de factores que una repetición de una experiencia o el empleo de cualquier otro instrumento puede modificar considerablemente el resultado obtenido la primera vez. La actitud científica más elemental nos obliga por tanto a preguntarnos si los valores observados son interpretables o si, por el contrario, son el resultado de fluctuaciones aleatorias introducidas por la propia medida. Este interrogante es particularmente necesario en la observación directa del comportamiento, donde los trabajos de investigación no dejan de poner en duda el valor de los registros observacionales.

Un instrumento fiable es aquel que tiene pocos errores de medida, y que muestra estabilidad, consistencia y dependencia en las puntuaciones individuales de las características evaluadas. Ahora bien, históricamente el estudio de la fiabilidad ha estado ligado al estudio de las diferencias individuales y por tanto casi restringido a las pruebas estandarizadas de inteligencia y personalidad. Sin embargo estas pruebas han ido poco a poco reemplazándose por observaciones de los individuos en situaciones naturales o cuasi-naturales, utilizando todas ellas observadores humanos para registrar el comportamiento de los individuos (Rowley, 1976; Smith & Teeter, 1982; Suen, Lee & Owen, in press).

Hay al menos tres formas de entender la fiabilidad de los registros procedentes de la observación directa del comportamiento (Berk, 1979; Blanco Villaseñor, 1983, 1986b, 1986c; Blanco Villaseñor y Anguera Argilaga, 1984; Medley & Mitzel, 1963; Mitchell, 1979).

1) En primer lugar, nos podemos referir a dos observadores que, registrando independientemente, codifican las conductas que ocurren. Este coeficiente de *concordancia* de los juicios de observadores, de acuerdo entre ellos, se refiere a las observaciones realizadas por diferentes observadores en el mismo momento —«coefficient of observer agreement»— (Anguera Argilaga, 1983, 1988).

2) En segundo lugar, una medida observacional puede considerarse como un caso especial de una prueba estandarizada (Mitchell, 1979), y en tal caso podemos utilizar las definiciones de fiabilidad de la teoría psicométrica clásica (López Feal, 1986) a través del coeficiente de correlación:

a) Fiabilidad *intraobservadores* (errores de comisión) o fiabilidad *interobservadores* (errores de omisión), es decir, obtener dos puntuaciones separadas de un mismo instrumento (o sesión de observación) (Blanco Villaseñor, 1983; Blanco Villaseñor y Anguera Argilaga, 1984). Se trata por tanto de verificar si coinciden los juicios cuantitativos emitidos por un observador en dos momentos diferentes sin interrupción temporal (mediante soportes audio y/o vídeo) o por dos

observadores en el mismo momento temporal (no necesariamente de forma simultánea si se dispone de soportes audio y/o vídeo) realizando un registro observacional.

b) *Equivalencia* (formas paralelas o equivalentes de sesiones de observación) u *homogeneidad* (dos mitades o partes de una misma sesión de observación), es decir obtener puntuaciones de dos instrumentos similares o de dos partes del mismo instrumento respectivamente.

c) *Constancia* (o estabilidad), es decir obtener puntuaciones del mismo instrumento en dos momentos diferentes, pero con una interrupción temporal.

Estos coeficientes de fiabilidad no son todo lo perfectos que podrían llegar a ser. Los coeficientes que utilizan dos puntuaciones de un mismo instrumento (intraobservadores e interobservadores) confunden el error aleatorio del individuo con las diferencias intra e interobservadores. Los coeficientes que utilizan puntuaciones de subdivisiones de una sesión de observación o de formas paralelas o similares de una sesión (homogeneidad y equivalencia) confunden el error aleatorio del sujeto con las diferencias entre las subdivisiones o formas. Y finalmente los coeficientes que utilizan puntuaciones del mismo instrumento administrado en dos ocasiones (constancia) confunden los errores de medida con los cambios reales que se producen en el comportamiento del sujeto en las dos ocasiones. Luego, estos métodos no permiten atribuir la variancia estimada a los observadores, a las formas diferentes, a las ocasiones, o no pueden considerar estas fuentes de error de forma simultánea. Es necesario, por tanto, una teoría multivariada que tenga en cuenta todas estas posibles fuentes de error, que será nuestra tercera forma de entender la fiabilidad de los registros observacionales.

3) Finalmente, una medición observacional puede presentar datos bajo la influencia de un cierto número de aspectos diferentes de una situación observacional (diferentes observadores, diferentes ocasiones, diferentes formas de registro, diferentes instrumentos de registro), incluyendo las diferencias individuales entre sujetos. Este tercer punto de vista es *la teoría de la generalizabilidad* desarrollada por Cronbach, Gleser, Nanda & Rajaratnam (1972), que asume que hay otras fuentes de variación además de las diferencias individuales y que permite integrar cada una de las fuentes de variación de los diferentes coeficientes de fiabilidad anteriores en una estructura global.

### *Aportaciones de la teoría de la generalizabilidad*

Cada vez se incrementa más el número de publicaciones psicológicas que advierten en sus textos de la importancia de la medición, tanto en investigación psicológica básica como aplicada. Sin embargo, la teoría de la generalizabilidad, aún a pesar de la advertencia de Brennan (1983, p. 12) «el modelo de medida mejor definido actualmente existente», ha sido subvalorada en la investigación psicológica y más aún dentro de la metodología observacional, donde mejor se adapta dadas sus particulares y peculiares condiciones de medida en diversas situaciones de un mismo sujeto.

El propósito del presente artículo es la revisión de los conceptos básicos

de la teoría de la generalizabilidad e ilustrar su utilización como un método de *diseñar, evaluar y optimizar* la dependencia-interdependencia («dependability»)<sup>1</sup> de las medidas psicológicas en general y procedentes de la observación directa del comportamiento en particular.

Como punto de partida, y para establecer una estructura de referencia, la teoría de la generalizabilidad siempre ha sido comparada con la teoría «clásica» de la fiabilidad. La distinción entre una y otra es más bien artificial, aunque tengan algunos elementos en común. Sin embargo las diferencias entre ambas teorías son más importantes que las similitudes. La teoría de la generalizabilidad aporta las siguientes ventajas:

1) La teoría de la generalizabilidad tiene su base en suposiciones menos restrictivas que la teoría clásica. Específicamente, el único supuesto a cumplir es el muestreo aleatorio de individuos y de condiciones de medida (observadores, situaciones, instrumentos de registro, ítems, entrevistadores, etc.). La teoría clásica con un cierto énfasis en las formas paralelas asume que las condiciones de medida son estrictamente equivalentes en contenido, media, variancia e intercorrelaciones (Cronbach, Rajaratnam & Gleser, 1963) y no tiene en cuenta el muestreo de los individuos (Cronbach et al., 1972, p. 127).

2) La teoría de la generalizabilidad reconoce explícitamente las múltiples fuentes de error de medida (observadores, situaciones, instrumentos de registro, etc.). Podemos estimar cada una de estas fuentes de error así como las diferentes interacciones entre ellas. Los efectos combinados de estas fuentes también podrán ser evaluados (Anguera Argilaga y Blanco Villaseñor, 1984). La fiabilidad clásica tiene en cuenta cada fuente de variación independientemente y no estima los efectos combinados.

3) La teoría clásica de la fiabilidad se ha desarrollado alrededor de los tests de inteligencia y personalidad asumiendo que una medida adquiere su máxima diferenciación a través de los individuos evaluados en un test (Cardinet, Tourneur & Allal, 1976, 1981). En la investigación observacional en particular y en otras situaciones de medida frecuentemente no se trata de diferenciar individuos sino más bien de diferenciar observadores, escalas, ocasiones, categorías de registro, grupos de individuos, etc. En palabras de Cardinet et al. (1976) esto significa asumir el principio de *simetría*, es decir que sucesivos objetos de medida pueden ser evaluados dentro de un mismo diseño. Mediante este principio, cada *faceta* o factor de un diseño puede ser seleccionado como objeto de estudio y en cada análisis de generalizabilidad de esta *faceta* puede ser considerada como instrumento de medida o condición de evaluación en el estudio de las otras facetas. La teoría clásica, por el contrario, sólo diferencia individuos. Imaginemos un índice de consistencia interna, como el  $\alpha$  de Cronbach, que nos muestre que un

1. Cronbach y sus colaboradores (1963, 1972) elaboraron primero totalmente la teoría de la generalizabilidad y acuñaron el término «dependability» (dependencia) de las mediciones. Sin embargo, este término no fue ni ha sido definido explícitamente a pesar de estar incluido en el propio título de su obra de 1972. Uno puede sospechar que el término acuñado implica que la teoría de la generalizabilidad es más general que la teoría clásica de la fiabilidad. De esta forma, podríamos definir dicho término entendiendo el hecho de que si un individuo o su comportamiento dependen de una medida particular sea ésta una representación precisa de todas las medidas similares que puedan realizarse en una situación particular de medida. Este concepto acompaña a los conceptos clásicos de consistencia interna y estabilidad, pero es evidente que también acompaña a otras fuentes de error de medida.

instrumento tiene baja fiabilidad de consistencia interna. Ello significa que la variancia atribuible a la interacción de individuos e ítems (error de variancia) es grande en comparación a la variancia atribuible a los individuos (variancia verdadera) y por tanto los ítems no diferencian bien a los individuos. Sin embargo, si nuestro interés se centra en diferenciar observadores, situaciones o instrumentos de registro, por ejemplo, dicho índice es poco significativo. Nuestro interés debemos buscarlo en el error de variancia referido a la variancia atribuible a la tasa media de los observadores en el conjunto de individuos (e ítems o categorías si las hubiere). Por tanto, consideramos que las medidas de fiabilidad clásica son casi siempre inapropiadas en contextos observacionales. Este hecho, además de los varios y diversos significados del término «fiabilidad», han encubierto probablemente la importancia de la evaluación de la fiabilidad en las investigaciones de observación directa del comportamiento. La teoría de la generalizabilidad reconoce explícitamente que los individuos no son siempre el objeto de medida y es capaz de estimar totalmente la dependencia de cada objeto de medida diferente (Rentz, 1987). Algunos autores han propuesto, históricamente hablando, adaptar los métodos clásicos (p. e. fiabilidad test-retest) a situaciones de observación directa en las que los objetos de medida no eran los individuos, pero dichos intentos en algunas ocasiones creemos que son vanos. Dichas situaciones pueden resolverse más fácilmente a través de la teoría de la generalizabilidad.

4) La teoría de la generalizabilidad es particularmente utilizada para el desarrollo y optimización de diseños de medida en estudios posteriores. Mediante el estudio sistemático de las diferentes fuentes de error podemos desarrollar un diseño de medida que reduzca el error total en los estudios siguientes. La teoría clásica tan solo nos aporta información acerca del número de ítems que tenemos que utilizar.

5) La dependencia de los valores medios obtenidos puede ser evaluada fácilmente a través de la teoría de la generalizabilidad. La investigación observacional está más interesada en puntuaciones medias que no en puntuaciones individuales.

6) Los estudios de generalizabilidad pueden ser llevados a cabo con planes muestrales estratificados y además aplicarse en el caso multivariado (Cronbach et al., 1972; Nussbaum, 1984; Shavelson & Webb, 1981). La teoría clásica no proporciona dichos análisis.

7) Los coeficientes de generalizabilidad, cuando se han optimizado de forma precisa, son claros y nunca ambiguos. Este tipo de diseños fuerzan al investigador a explicitar el universo al cual se desea generalizar.

Estos puntos quedarán mucho más claros cuando hayamos acabado de examinar la teoría de la generalizabilidad. En los siguientes apartados revisaremos los conceptos básicos de la teoría de la generalizabilidad, presentaremos un método simplificado para el cálculo de los coeficientes de generalizabilidad (concretamente el propuesto por Cardinet et al., 1976, y no el especificado en la obra original de Cronbach et al., 1972) e ilustraremos con un ejemplo la aplicación de la teoría.

### *Teoría de la generalizabilidad: conceptos básicos*

La generalizabilidad es una teoría de los errores multifaceta de una medición conductual (Cronbach et al., 1972). El objetivo de la teoría es desglosar, en cualquier tipo de medición, la variabilidad real de la variabilidad del error. Para que se cumpla, la teoría necesita de los componentes del análisis de la variancia; las variancias de las facetas (variables, factores) tales como individuos, observadores, sesiones, ocasiones de medida. El eje central de la teoría de la generalizabilidad, por tanto, se encuentra en los componentes de variancia, dado que su magnitud nos aporta información sobre las fuentes de error que están afectando una medición conductual.

La filosofía básica que subyace a la teoría de la generalizabilidad es que «un investigador se pregunta acerca de la precisión o fiabilidad de una medición, dado que desea generalizar de observaciones reales a cualquier tipo de observaciones a las que éstas pertenezcan» (Cronbach, Rajaratnam & Gleser, 1963, p. 144). En un estudio observacional, esto es un universo al que el observador desea generalizar. El conjunto de condiciones de medida sobre las que el investigador-observador generaliza es el universo de generalización. Éste diferirá de acuerdo a los propósitos de la investigación. Pero en todo caso es importante que el investigador-observador defina claramente el universo especificando las condiciones de medida sobre las que intenta generalizar.

Las nociones clásicas de puntuación verdadera y del error van a ser reemplazadas por conceptos más en armonía con desarrollos estadísticos modernos (Martínez Arias, 1981). Sea  $x_{(ic)}$  el valor observado del individuo (i) en la categoría (c). Imaginemos que el sujeto (i) sea evaluado a través de todas las categorías (c), que se refieren a la misma competencia, y que calculamos la media de los  $x_{(ic)}$  obtenidos. Obtendremos una *puntuación universo* para el individuo (i), que designaremos por  $\mu(i)$ , que es un indicador del grado de posesión del sujeto de la competencia evaluada. Si por el contrario, todos los sujetos de una misma población son evaluados en una misma categoría (c), obtendremos una media  $\mu(c)$  que constituirá la *puntuación universo* ligada a la categoría (c) para todas las personas de la población. Es decir, habríamos calculado la ocurrencia de esta categoría a partir de los individuos de la población. Si calculamos la media para todas las  $\mu(c)$  o para todas las  $\mu(i)$  obtendremos la media general  $\mu$  de todas las categorías (c) del universo (que miden dicha competencia) aplicadas a todos los sujetos de la población.

El concepto de *puntuación universo* explica el hecho de que lo que interpreta una medida es la estimación, a partir de una muestra de datos observados, de un valor teórico inobservable. Se intenta conocer la media de todos los valores del individuo (i) que se obtendrían si se efectuasen las observaciones en todas las condiciones posibles. Luego, el investigador-observador utiliza la puntuación observada, o una función de la puntuación observada para poder estimar el valor de la puntuación universo. Así se generaliza de la muestra a la población. El problema de la fiabilidad es por tanto el de la precisión de esta generalización (o de su «generalizabilidad» en terminología de Cronbach et al., 1972, p. 15).

La generalizabilidad es así el grado por el cual podemos generalizar un resultado obtenido en unas condiciones particulares a un valor teórico buscado.

El *universo de generalización* es el conjunto de condiciones a las que se quiere generalizar los resultados observados en sus condiciones particulares. Ello resulta eventualmente de la elección de un subconjunto de condiciones «admisibles» en el conjunto original de todas las condiciones posibles (Cardinet & Tourneur, 1985).

El término *faceta* lo introducen Cronbach et al. (1972) con el fin de designar cada una de las características de la situación de medida que es susceptible de ser modificada de una observación a otra y que puede hacer variar en consecuencia el valor del resultado obtenido. Por ejemplo, las categorías en las que pueden ser evaluados los sujetos van a variar de un sujeto a otro según el observador y por tanto constituyen una fuente de variación importante. Así, hablaremos de la faceta «categorías».

Cada una de las manifestaciones posibles de una faceta (cada observador, cada individuo, cada categoría, cada método de registro de datos, cada instrumento, etc.), ya que cada elemento del conjunto constituye una faceta, será designada como un *nivel* de la faceta.

Dos o más facetas serán *cruzadas* si se dispone al menos de un dato para cada combinación de niveles de una faceta con los niveles de la otra (una cruz simbolizará la relación cruzada entre facetas). Una faceta estará *anidada* en otra si los niveles de la primera faceta (anidada) son diferentes de un nivel en la otra segunda faceta (anidante). La simbolización de un doble punto (:) separa la faceta anidada (a la izquierda) de la anidante (a la derecha).

Una *faceta es aleatoria* si una muestra aleatoria simple de *niveles observados* ha sido extraída de un conjunto infinito (o hipotéticamente infinito) de *niveles admisibles*. Los niveles *admisibles* de cada faceta corresponden al número posible de objetos de estudio y de instrumentos de medida. El número de niveles observados en la muestra será simbolizado por  $n$  seguido de la letra de la faceta. El número de niveles admisibles en la población o en el universo estará simbolizado por  $N$ . En el caso de que una faceta sea *fija*  $N_{(i)} = n_{(i)}$ . Una faceta es *fija* si los niveles observados agotan los niveles admisibles. Habría que considerar un tercer caso, intermedio entre los dos anteriores, el de las facetas que están constituidas por un muestreo aleatorio a partir de una población con universo finito de niveles (Cardinet & Tourneur, 1985): facetas aleatorias *finitas*.

Los objetos de medida admisibles constituyen la población objeto de estudio y los instrumentos de medida (las condiciones de evaluación en terminología de Cronbach) constituyen el universo de generalización. Los primeros se sitúan en el aspecto de la *diferenciación*, ya que la variancia verdadera proviene de las diferencias entre objetos de estudio. Los segundos se sitúan en el aspecto de *instrumentación* (o generalización), puesto que las condiciones de medida son como los instrumentos o medios de esta medida. Es decir, estimación de la variancia verdadera debida a las diferencias entre los objetos de medida, y estimación de la variancia de error debida a elección de los instrumentos utilizados en la medida, respectivamente.

Todo evaluador tiene en mente un universo al que propone generalizar sus

observaciones. Este universo define las fuentes de variación que le interesan y que va a tener en cuenta. Con este fin, debe estimar todos los componentes de variancia de las observaciones en un estudio previo, al que denominaremos *estudio de generalizabilidad* (G). Luego elegirá un nuevo plan de observación que tratará de minimizar los componentes «parásitos», no deseados, de la variancia de las puntuaciones: ello será objeto de un *estudio de decisión* (D), que aprovechará las informaciones obtenidas en el estudio G.

### *Teoría de la generalizabilidad: bases teóricas*

Utilizaremos un sencillo diseño para presentar las bases teóricas de un análisis de generalizabilidad. El lector familiarizado con el análisis de la variancia (AVAR) reconocerá que la presentación es fácilmente extensible a diseños mucho más complejos (Arnau, 1978, 1981). En principio, esta presentación teórica está basada en el trabajo de Shavelson & Webb (1981), para posteriormente ir adaptándola a los de Cardinet, Tourneur & Allal (1976, 1981) y de Cardinet & Tourneur (1985). Asumiremos en este diseño simple, aunque sólo sea por simplicidad, que los individuos son objeto de medida. Esta suposición no es necesaria, puesto que por el principio de simetría, del que hablamos anteriormente, cualquier objeto de medida puede ser sustituido por la faceta individuos que utilizamos aquí.

El análisis de la generalizabilidad presupone que la puntuación de interés es una puntuación media (o suma) a través de diferentes muestras de condiciones de medida. Raras veces son las que estamos interesados en la respuesta de un individuo a una categoría o ítem en concreto, sino que más bien nos centramos en la puntuación media (o suma) de un individuo a través de diferentes muestras de todas las categorías o ítems en el universo de generalización. Imaginemos un diseño simple individuo (i) por categoría (c), con generalización a todas las categorías admisibles. Se supone que las condiciones de ambas facetas han sido muestreadas al azar de la población de individuos y del universo de las categorías o ítems, respectivamente, y que el diseño es cruzado, es decir, cada individuo ha sido evaluado en cada una de las categorías (en la teoría de la generalizabilidad el diseño cruzado se representa con letras mayúsculas y con el signo de la multiplicación I×C).

La puntuación de un determinado individuo en una determinada categoría ( $x_{ic}$ ) se representa como sigue:

$$X_{ic} = \mu + (\mu_i - \mu) + (\mu_c - \mu) + (X_{ic} - \mu_i - \mu_c + \mu), \quad (1)$$

donde

$$\begin{aligned} \mu &= \text{la media general} \\ \mu_i - \mu &= \text{el efecto del individuo } i \\ \mu_c - \mu &= \text{el efecto de la categoría } c \\ X_{ic} - \mu_i - \mu_c + \mu &= \text{residual} \end{aligned}$$

Cada componente de una puntuación (excepto para la media general) tie-

ne una distribución con media cero. La variancia de la distribución de  $\mu_i - \mu$  es  $\Sigma(\mu_i - \mu)^2 = \sigma_i^2$ . De igual forma, la variancia del efecto de las categorías es  $\Sigma(\mu_c - \mu)^2 = \sigma_c^2$  y la variancia de residuales es  $\Sigma(X_{ic} - \mu_i - \mu_c - \mu)^2 = \sigma_{ic,e}^2$ . El subíndice [ic,e] indica que la variancia de interacción y la variancia del error residual están confundidas en los diseños simples de una observación por casilla. Generalmente, se escribe simplemente  $\sigma_{ic}^2$  en lugar de  $\sigma_{ic,e}^2$ . La variancia de todas las puntuaciones observadas ( $\sigma_{X_{ic}}^2$ ) es  $\Sigma(X_{ic} - \mu)^2$  y es igual a la suma de todos los componentes de variancia.

$$\sigma_{X_{ic}}^2 = \sigma_i^2 + \sigma_c^2 + \sigma_{ic,e}^2 \quad (2)$$

La teoría de la generalizabilidad se centra precisamente sobre los componentes de variancia, dado que su magnitud nos aporta cierta información sobre las fuentes de error que están afectando a una medición. Estos componentes de variancia vendrán determinados por un estudio de generalizabilidad: «El instrumento desarrollado, que se utiliza en un estudio G como guía del usuario de dicho instrumento, tratará sistemáticamente en el diseño objeto de estudio, las facetas que posiblemente entrarán a formar parte de generalizaciones de varios estudios» (Cronbach et al., 1972, p. 21). Los componentes de variancia se estiman como tradicionalmente se hace en el AVAR de datos aleatorios, aunque hay otros modos más complejos de estimación. Tres métodos alternativos, aunque similares, pueden encontrarse en tres trabajos diferentes, Brennan (1983), Cardinet, Tourner & Allal (1976) y Marcoulides (1989). La forma de estimar los componentes de variancia es igualando los cuadrados medios observados del AVAR a los valores esperados y resolviendo el conjunto de ecuaciones lineales. Para ilustrar una estimación de un diseño simple, consideremos el ejemplo anterior I x C (Tabla 1).

TABLA 1. ESTIMACIÓN MEDIANTE AVAR DE LOS COMPONENTES DE VARIANCIA

Fuente de Variación	CM	Cuadrado Medio Esperado (CME)
Individuos (i)	CM <sub>i</sub>	$\sigma_{ic,e}^2 + n_c \sigma_i^2$
Categorías (c)	CM <sub>c</sub>	$\sigma_{ic,e}^2 + n_i \sigma_c^2$
Errores (ic,e)	CM <sub>ic,e</sub>	$\sigma_{ic,e}^2$

Los valores estimados para cada componente de variancia se obtienen del AVAR como sigue:

$$\sigma_i^2 = [CM_i - CM_{ic,e}] / n_c$$

$$\sigma_c^2 = [CM_c - CM_{ic,e}] / n_i$$

$$\sigma_{ic,e}^2 = CM_{ic,e}$$

Si suponemos que los individuos constituyen la faceta de diferenciación

y las categorías la faceta de generalización,  $\sigma^2$ , es análoga a la variancia verdadera en la teoría clásica, siendo la variancia de puntuaciones universo en la teoría de la generalizabilidad.  $\sigma^2_c$  y  $\sigma^2_{ic}$  contribuyen a la variancia del error.

La teoría de la generalizabilidad trata esencialmente la descomposición de la variancia observada en componentes de variancia y obtiene información analizando dichos componentes, particularmente en lo que respecta a la contribución del error en un determinado diseño. El análisis de los componentes informa sobre qué facetas contribuyen con más error, para ser modificadas posteriormente en los sucesivos diseños. Un índice global de la variancia de puntuaciones universo relativa a la variancia del error es el coeficiente de generalizabilidad.

Un coeficiente de generalizabilidad es un coeficiente de correlación intraclassa similar en forma al tradicional coeficiente de la fiabilidad clásica. El coeficiente de generalizabilidad ( $E\hat{\rho}^2$ ) se define como la proporción de variancia observada que es atribuible a la puntuación universo, es decir, es la razón entre el valor esperado de la variancia de puntuaciones universo ( $\sigma^2$ ) y el valor esperado de la variancia de puntuaciones observadas ( $E\sigma^2_x$ )

$$E\hat{\rho}^2 = \sigma^2_{\tau} / E\sigma^2_x \quad (3)$$

o bien, dado que  $E\sigma^2_x = \sigma^2_{\tau} + \sigma^2_{\delta}$  es la variancia de error relativo (que ya definiremos posteriormente),

$$E\hat{\rho}^2 = \sigma^2_{\tau} / [\sigma^2_{\tau} + \sigma^2_{\delta}] \quad (4)$$

El lector notará una similitud entre la definición del coeficiente de generalizabilidad y la del coeficiente de la fiabilidad clásica. Ambas implican la razón entre la variancia de puntuaciones universo (o verdadera) y la variancia de puntuaciones observadas. Sin embargo, hay dos importantes diferencias. La primera es que la teoría de la generalizabilidad admite múltiples fuentes de variación en las puntuaciones observadas mientras que la teoría clásica sólo admite una (ítems u ocasiones, etc.) en cualquier coeficiente de fiabilidad y define múltiples coeficientes de fiabilidad (test-retest, consistencia interna, etc.). En segundo lugar, dado que la teoría de la generalizabilidad presupone un muestreo aleatorio de los niveles de las facetas, el coeficiente de generalizabilidad está basado en la variancia de puntuaciones observadas «esperadas» mientras que los coeficientes clásicos no. Luego, el coeficiente de generalizabilidad nos permite generalizar los valores del coeficiente a otras situaciones de medida paralelas (aleatoriamente). Decimos que dos medidas son paralelas aleatoriamente si ambas han sido muestreadas aleatoriamente del mismo conjunto de condiciones en el universo de generalización. Este tipo de medidas paralelas aleatoriamente no necesitan cumplir las propiedades paralelas restrictivas de la teoría clásica: igual media, igual variancia, e igual intercorrelación (Brennan, 1983).

Podemos obtener una estimación del coeficiente de generalizabilidad ( $E\hat{\rho}^2$ ) a través de estimaciones muestrales de los parámetros de la ecuación anterior.

$$E\hat{\rho}^2 = \hat{\sigma}^2_{\tau} / [\hat{\sigma}^2_{\tau} + \hat{\sigma}^2_{\delta}] \quad (5)$$

$E\hat{\rho}^2$  es sesgado, pero es un estimador consistente de  $E\rho^2$  (Shavelson & Webb, 1981; Shavelson, Webb & Rowley, 1989).

En la teoría de la generalizabilidad se distinguen tres tipos de error, dos de los cuales son comúnmente utilizados y los mencionaremos aquí<sup>2</sup>. El tipo de error que se utiliza dependerá del tipo de interpretación de las puntuaciones que el investigador quiera hacer. La *variancia de error relativo*, representada por  $\sigma^2_{\delta}$ , es útil cuando se quiere hacer una interpretación relativa de las puntuaciones. Por ejemplo, se quiere evaluar a 20 individuos en una selección de personal, *clasi-ficándose* para pruebas posteriores tan solo *los dos* que obtengan la puntuación más alta. Esta decisión es relativa y  $\sigma^2_{\delta}$  es el término de error apropiado a utilizar. Es decir, las medidas relativas son las posiciones relativas de los resultados y en todos los casos la escala objetiva del instrumento de medida no agota toda la información que podemos obtener de un resultado. El error relativo «puede ser apropiado utilizarlo cuando se quiere determinar si uno de dos individuos evaluados bajo las mismas condiciones tiene una puntuación uníversono significativamente más alta que el otro. Es relevante si una decisión depende de una clasificación» (Cronbach et al., 1972, p. 355). Muchas de las decisiones en psicología y más concretamente en metodología observacional son decisiones relativas, por lo que el error relativo es el apropiado en la mayoría de los casos.

La variancia de error absoluto, representada por  $\sigma^2_{\Delta}$ , es apropiada cuando se quieren interpretar las puntuaciones en un sentido absoluto. Por ejemplo, se clasificarían para pruebas posteriores los individuos que excedieran una mínima puntuación. Esta decisión es absoluta y el error absoluto  $\sigma^2_{\Delta}$  es el apropiado. Es decir, hablamos de medida absoluta cuando queremos situar una magnitud con respecto a una escala, en la que los escalones han sido definidos a priori, antes de que se hayan efectuado las observaciones. El error absoluto «indica cómo medidas distantes tienen probablemente su punto de partida de sus valores 'verdaderos'; por ejemplo, de la puntuación uníversono del individuo» (Cronbach et al., 1972, p. 355)<sup>3</sup>. Si queremos hacer tales interpretaciones debemos modificar el coeficiente de generalizabilidad simplemente sustituyendo  $\sigma^2_{\delta}$  por  $\sigma^2_{\Delta}$  en las dos ecuaciones anteriores.

La diferencia entre  $\sigma^2_{\delta}$  y  $\sigma^2_{\Delta}$  es que la  $\sigma^2_{\delta}$  no incluye las fuentes de variancia común a cada individuo (objeto de medida) mientras que la  $\sigma^2_{\Delta}$  sí. Por tanto, la  $\sigma^2_{\Delta}$  tendrá valores como mínimo iguales a  $\sigma^2_{\delta}$ , aunque casi siempre los valores serán más altos. Es interesante hacer notar que en un diseño simple individuos por categorías,  $\sigma^2_{\delta}/[\sigma^2_{\delta} + \sigma^2_{\Delta}]$  es algebraicamente (pero no conceptualmente) idéntico al  $\alpha$  de Cronbach (Brennan, 1983).

La discusión que se ha llevado a cabo sobre las bases teóricas de los coeficientes de generalizabilidad ha sido necesariamente breve, pero el lector interesado podrá consultar las obras de Cronbach et al. (1972) y Brennan (1983). En el

2. El tercer tipo es apropiado cuando una decisión está basada en una estimación mediante regresión de la puntuación uníversono. Este tipo de error casi nunca se ha aplicado en la literatura científica y probablemente tiene aplicaciones limitadas en metodología observacional. Se puede consultar la obra de Cronbach et al. (1972) o el trabajo de Shavelson & Webb (1981) que incluye un breve comentario.

3. El error absoluto ha sido muy utilizado en la interpretación de las puntuaciones de los tests referidos al criterio. Más información concreta sobre este aspecto puede encontrarse en los trabajos de Brennan (1983), Kane & Brennan (1980) y Shavelson & Webb (1981, p. 145).

caso de una breve revisión, consultar Shavelson & Webb (1981) y Shavelson, Webb & Rowley (1989).

### *Análisis de la generalizabilidad: procedimiento*

Los complejos algoritmos de cálculo en su formulación original propuesta por Cronbach et al. (1972) han sido simplificados y desarrollados especialmente en los trabajos de Brennan (1980, 1983), Cardinet & Tourneur (1985), y Cardinet, Tourneur & Allal (1981). Los procedimientos y fases descritas por Cardinet, Tourneur & Allal (1981) son los que describiremos aquí, ya que son aplicables a un rango muy amplio de diferentes diseños de medida, y además relativamente fáciles de aplicar.

Para ilustrar nuestro análisis de generalizabilidad utilizaremos tres diseños (resumidos en la Tabla 2). En los tres mostraremos los cálculos de las variancias y de los coeficientes de generalizabilidad, así como los valores numéricos obtenidos en cada uno de ellos.

El primero de ellos es un diseño simple cruzado con tres facetas. En este hipotético ejemplo que proponemos, 5 observadores han de valorar a 4 individuos en una escala de estimación. Cada individuo deberá ser evaluado en una escala que contiene 3 categorías, y cada observador podrá evaluar en la escala hasta una puntuación de 5. Dicho plan de observación comprenderá tres facetas cruzadas: los observadores (O, que son 5), los individuos (I, un total de 4) y las categorías (C, que son 3). Las tres facetas son cruzadas, ya que la respuesta de cada individuo a cada categoría deberá ser anotada por el observador. El plan se simbolizará por  $O \times I \times C$  y los valores figurarán en una tabla de doble entrada que contiene ( $5 \times 4 = 20$ ) casos y 3 datos para cada caso. El investigador considera que los observadores y las categorías son fuentes de error y por tanto intenta diferenciar a los individuos sobre la base de las puntuaciones dadas por los observadores a cada una de las categorías. Así, los individuos (I) serán la faceta de diferenciación, mientras que observadores (O) y categorías (C) constituirán las facetas de generalización. De esta forma, se trata de ver si podemos generalizar las puntuaciones al universo de observadores y categorías. Todas las facetas serán consideradas aleatorias.

El hecho de determinar si los observadores y/o las categorías son aleatorios, finitos o fijos puede plantear dificultades conceptuales. En el caso de las categorías suelen considerarse generalmente aleatorias extraídas de un universo de categorías, dado el rango infinito de valores que distancia a una categoría de otra; sin embargo, las categorías en un instrumento dado agotan el conjunto de posibles categorías. En la práctica se suelen considerar facetas aleatorias, pero el hecho de considerarlas fijas se puede determinar fácilmente en un análisis de generalizabilidad. En cuanto a los observadores, casi siempre quedan reducidos al número de profesores o de investigadores, por lo que casi siempre suele considerarse faceta fija, aunque nada impide que se pueda considerar finita o aleatoria. El tema de las facetas finitas tan solo es considerado en el trabajo de Cardinet & Tourneur (1985), y en el caso de utilizarlas exige una serie de correcciones en el cálculo de las variancias de diferenciación y de error absoluto y relativo.

TABLA 2. TRES DISEÑOS HIPOTÉTICOS PARA ILUSTRAR EL ANÁLISIS DE LA GENERALIZABILIDAD

<b>Plan de observación</b>	$O \times I \times C$	$(J : T) \times F \times E$	$(J : T) \times F \times E$
	O=5 I=4 C=3	J=5 T=2 F=6 E=4	I=5 T=2 F=6 E=4
<b>Anidaciones</b>	Ninguna	J anidado en T	J anidado en T
<b>Plan de estimación</b>	Aleatoria O, I, C	Aleatoria J, T, F, E	Aleatoria, J, F, E
	Finita—	Finita—	Finita—
	Fija—	Fija—	Fija T
<b>Fuentes de variación</b>	O,I,C,IO,OC,IC,OIC	T,F,E,J,T,TF,TE,FE, FJT,EJ,T,TFE,FEJT	T,F,E,J,T,TF,TE,FE FJT,EJ,T,TFE,FEJT
<b>Plan de medida</b>	(I/—/—/OC)	(J,T/—/—/F,E)	(—/T/—/J,F,E)
<b>Faceta(s) de diferenciación</b>	Individuos	Jueces, Tomas	Tomas
<b>Faceta(s) de instrumentación</b>	Observadores, Categorías	Fotografías, Escalas	Jueces, Fotografías, Escalas
<b>Generalizabilidad</b>			
<b>Puntuación Universo:</b>			
<b>Componentes de variancia <math>\hat{\sigma}_r^2</math></b>	$\hat{\sigma}_r^2 = 0,0948$	$\hat{\sigma}_r^2 + \hat{\sigma}_{j:t}^2 = 1,4154$	$\hat{\sigma}_r^2 = 0$
<b>Error absoluto:</b>			
<b>Componentes de variancia <math>\hat{\sigma}_A^2</math></b>	$\hat{\sigma}_{oi}^2/n_o + \hat{\sigma}_{ic}^2/n_c + \hat{\sigma}_{oi}^2/n_o + \hat{\sigma}_{oc}^2/n_o n_c +$ $+ \hat{\sigma}_{ic}^2/n_c + \hat{\sigma}_{oic}^2/n_o n_c = (0,0153/5) +$ $+ (4,5549/3) + (0,1649/5) +$ $+ (0,0312/15) + (0,7733/3) +$ $+ (0,4615/15) = 1,8449$	$\hat{\sigma}_{jt}^2/n_f + \hat{\sigma}_{te}^2/n_e + \hat{\sigma}_{jt}^2/n_f + \hat{\sigma}_{te}^2/n_e +$ $+ \hat{\sigma}_{fj:t}^2/n_f n_e + \hat{\sigma}_{te}^2/n_f n_e + \hat{\sigma}_{ej:t}^2/n_e +$ $+ \hat{\sigma}_{fje:t}^2/n_f n_e + \hat{\sigma}_{fej:t}^2/n_f n_e = 0,93$	$\hat{\sigma}_{j:t}^2/n_j + \hat{\sigma}_{f:e}^2/n_e + \hat{\sigma}_{j:t}^2/n_j + \hat{\sigma}_{f:e}^2/n_j +$ $+ \hat{\sigma}_{ej:t}^2/n_e n_j + \hat{\sigma}_{fje:t}^2/n_f n_e +$ $+ \hat{\sigma}_{fej:t}^2/n_f n_e n_j = 0,4724$
<b>Error relativo:</b>			
<b>Componentes de variancia <math>\hat{\sigma}_D^2</math></b>	$\hat{\sigma}_{oi}^2/n_o + \hat{\sigma}_{ic}^2/n_c + \hat{\sigma}_{oic}^2/n_o n_c =$ $= (0,1649/5) + (0,7733/3) + (0,4615/15) =$ $= 0,3215$	$\hat{\sigma}_{jt}^2/n_f + \hat{\sigma}_{te}^2/n_e + \hat{\sigma}_{fj:t}^2/n_f + \hat{\sigma}_{te}^2/n_e +$ $+ \hat{\sigma}_{fje:t}^2/n_f n_e + \hat{\sigma}_{fej:t}^2/n_f n_e = 0,9225$	$\hat{\sigma}_{j:t}^2/n_j + \hat{\sigma}_{f:e}^2/n_e + \hat{\sigma}_{fje:t}^2/n_f n_j +$ $+ \hat{\sigma}_{fej:t}^2/n_e n_j + \hat{\sigma}_{fje:t}^2/n_f n_e + \hat{\sigma}_{fej:t}^2/n_f n_e n_j =$ $= 0,4704$
<b>Coefficiente generalizabilidad:</b>			
<b>Absoluto</b>	$Eg_{A(i)}^2 = 0,0489$	$Eg_{A(j,i)}^2 = 0,6035$	$Eg_{A(i)}^2 = 0$
<b>Relativo</b>	$Eg_{D(i)}^2 = 0,2287$	$Eg_{D(j,i)}^2 = 0,6054$	$Eg_{D(i)}^2 = 0$
<b>Plan de optimización</b>	El Plan de Optimización se comentará en los siguientes apartados.		

El segundo diseño es otro hipotético ejemplo que ilustra un plan más complejo de 4 facetas y alguna de ellas anidada. Es una supuesta investigación sobre mapas cognitivos donde diferentes expertos en urbanismo de una determinada ciudad tratan de valorar diferentes fotografías de lugares públicos (Blanco Villaseñor, 1986a), evaluando cada una de las fotografías en 4 escalas de estimación diferentes (bienestar social, satisfacción, arquitectura, historia) con el fin de especificar si el lugar es poco conocido o no tiene interés social y público o por el contrario si el lugar es muy reconocido por su interés social y público. Las fotografías se presentarán en dos ocasiones distintas, lo que implicará que la segunda toma sea desde diferente perspectiva que la primera (por ejemplo, fotografía tomada desde el lado norte y fotografía tomada desde el lado sur). Dicho plan de observación comprenderá 4 facetas: los expertos en urbanismo que denominaremos jueces (J, que son 5), las diferentes fotografías de lugares públicos (F, un total de 6), las distintas escalas de estimación (E, que son 4) y las dos perspectivas de la toma de la fotografía (T, igual a 2).

Es un plan mixto anidado-cruzado en el que los jueces están anidados en las tomas de la fotografía, simbolizando este plan por  $(J:T) \times F \times E$ . El intento es el de diferenciar a los jueces, ya que ellos constituirán una faceta de diferenciación. El estudio G también administrará cada una de las 4 escalas a cada fotografía y por tanto en diferentes ocasiones con el fin de estimar la generalizabilidad a través de diferentes escalas y en diferentes ocasiones. De esta forma, fotografías y escalas conformarán las facetas de generalización, ambas consideradas aleatorias.

El investigador espera que las tomas puedan contribuir a la variación en las puntuaciones de los jueces. Teóricamente, las tomas han sido seleccionadas aleatoriamente y en cada toma también los jueces han sido seleccionados al azar. Por lo tanto, las tomas estarán incluidas como faceta de diferenciación y los jueces quedarán anidados dentro de las tomas.

En general, las facetas en clasificación y/o estratificación deberían de incluirse como facetas de diferenciación con el fin de determinar si estas facetas afectan la variancia de los objetos que van a ser diferenciados (Rentz, 1987). Si los componentes de variancia de estas facetas son pequeños, deben ser eliminadas en los estudios posteriores de optimización (Cardinet, Tourneur & Allal, 1981). Si el objeto de medida está estratificado respecto a subpoblaciones fijas, sería lógico realizar análisis por separado para cada nivel de la subpoblación además del análisis global. Por ejemplo, si en el segundo diseño las tomas fueran consideradas fijas, se realizaría un análisis diferente para cada toma. En el caso de que los componentes de variancia difieran sustancialmente en las diferentes subpoblaciones (tomas), la medición estará sesgada y por tanto se tomarán decisiones por separado de cada subpoblación para eliminar el sesgo. Sobre este tema se pueden consultar los trabajos de Brennan (1983, pp. 93-97), Cardinet, Tourneur & Allal (1981, pp. 201-202) y Shavelson & Webb (1981, pp. 147-148).

El tercer diseño que proponemos es hipotético y está basado en los mismos datos que el diseño 2, pero el investigador tiene unos objetivos diferentes. En este caso, está interesado en la variación de los jueces dentro de cada una de las tomas y por tanto necesita una medida de las tomas. En cada toma, los

jueces serán seleccionados al azar. De esta forma, las tomas serán el objeto de estudio puesto que la toma (T) es la faceta de diferenciación. Los jueces serán aquí faceta de generalización, mientras que en el diseño 2 eran de diferenciación. Con estos dos diseños pretendemos ilustrar el hecho de que el universo de generalización y la generalizabilidad de las escalas utilizadas pueden diferir entre investigaciones, dependiendo del tipo de interpretación que hagamos de las puntuaciones.

En este último diseño, se presupone que se incluyen todas las tomas posibles, ya que las tomas (T) es una faceta fija de diferenciación. Jueces, fotografías y escalas serán facetas de generalización, todas aleatorias.

La tabla 3 resume los procedimientos de cálculo para llevar a cabo un análisis de generalizabilidad, según una adaptación de la presentada por Cardinet et al. (1981). Dichos procedimientos implican tres fases (1) las técnicas tradicionales del AVAR, (2) las técnicas de un análisis de generalizabilidad y (3) optimización del diseño inicial basada en los resultados de las fases 1 y 2. A las 2 primeras fases Cronbach et al. (1972) les denomina estudio G (generalizabilidad) y a la tercera estudio D (decisión). Cardinet et al. (1981) denominan Plan de Observación a los puntos 1 y 2, Plan de Estimación a los puntos 3 y 4, Plan de Medida a los puntos 5-11 y Plan de Optimización a la Fase 3.

TABLA 3. PROCEDIMIENTOS DE CÁLCULO DE UN ANÁLISIS DE GENERALIZABILIDAD

#### FASE 1. AVAR

1. Identificación fuentes de variación ( $\alpha$ ).  
Construir la tabla del AVAR especificando las fuentes de variación de todas las facetas de diferenciación y de generalización, así como de todas sus interacciones. El subíndice que describe un efecto se especificará de la siguiente manera (subíndice/s primario/s): (1<sup>o</sup> subíndice anidado):...(enésimo subíndice anidado). Por ejemplo, en el diseño 2, si los jueces estuvieran anidados en las tomas y éstas a su vez anidadas en las escalas, escribiríamos J:T:E.
2. Calcular el cuadrado medio (CM) de cada una de las fuentes de variación:  $CM_{(a)}$ .
3. Componentes de variancia de un modelo totalmente aleatorio  $\sigma^2_{(a)}$ . Calcular la estimación de los componentes de variancia de cada efecto, suponiendo que todos son aleatorios.
4. Componentes de variancia de un modelo mixto (aleatorio-finito-fijo):  $\sigma^2_{(a/M)}$ . Si el diseño incluyera facetas finitas o fijas, ajustar los correspondientes componentes del modelo mixto. Ello significa que utilizaremos dichas estimaciones en los pasos que siguen.

#### FASE 2. ANÁLISIS DE GENERALIZABILIDAD

5. Diseño/s de medida.  
Seleccionar los diseños que serán utilizados en los pasos posteriores con el siguiente esquema: (finitas o infinitas/fijas/fijas/finitas o infinitas)  

	diferenciación		generalización	
--	----------------	--	----------------	--
6. Control de coherencia.  
Verificar que ninguna faceta de diferenciación esté anidada en una de generalización.
7. Variancia activa (Presencia de facetas fijas).
  - a) Eliminar todos los componentes de variancia que contengan facetas fijas de generalización en su subíndice primario. Todos los componentes eliminados conformarán la variancia pasiva.
  - b) Ajustar las facetas fijas y/o finitas de diferenciación, multiplicando todos los componentes que tengan en su subíndice una faceta fija y/o finita de diferenciación por el índice correc-

tor  $(n_f-1)/n_f$ , siendo  $n_f$  el número de niveles admisibles de la faceta cuyo número de niveles es finito o fijo. Los componentes ajustados se denominarán «esperanza de variancia de  $\alpha$ » y se simbolizarán  $E^2_{(\alpha)}$ .

8. Variancia de diferenciación ( $\sigma^2_{\tau}$ ).  
Es la suma de todos los componentes de la variancia activa que *sólo* contienen facetas de diferenciación en su subíndice primario.
9. Variancia de error absoluto ( $\sigma^2_{\delta}$ ).  
a) Suma de todos los componentes restantes (excepto los que ya han entrado a formar parte de la  $\sigma^2_{\tau}$ ), dividiendo cada uno por el producto del número de *niveles* observados de las facetas de instrumentación que aparecen en su subíndice *total*.  
b) Para cada faceta de instrumentación aleatoria finita ( $\infty > N > n$ ) que aparezca en el subíndice *primario* de un componente, multiplicar el término obtenido en la etapa precedente por el índice corrector del universo finito correspondiente:  $[N_{(f)} - n_{(f)}] / [N_{(f)} - 1]$ .
10. Variancia de error relativo ( $\sigma^2_{\delta}$ ).  
Se obtiene sumando, con sus coeficientes (incluso los correctores), todos los componentes de la variancia de error absoluto (paso 9) que contienen al menos una faceta de diferenciación en su subíndice *total*.
11. Coeficiente/s de generalizabilidad  $Eg^2_{(A)}$  o  $Eg^2_{(d)}$ .  
Se calcula dividiendo la variancia de diferenciación por la suma de la variancia de diferenciación y de la variancia de error (relativa o absoluta, según la puntuación que se considere).

### FASE 3. MODIFICACIÓN Y OPTIMIZACIÓN

12. Dependiendo de los resultados obtenidos en los análisis precedentes, definir las modificaciones que se van a realizar en el diseño, con el fin de disminuir el error o reducir costos. Para ello se necesita repetir los pasos 3, 4, 6 y 7.
13. Calcular de nuevo los pasos 8 a 11 para obtener  $\sigma^2_{\tau}$ ,  $\sigma^2_{\delta}$ ,  $\sigma^2_{\delta}$ , y así obtener  $Eg^2$  optimizado.

\* Adaptada de Cardinet, Tourneur & Allal (1981) y Cardinet & Tourneur (1985).

El objetivo de la primera fase es la estimación de los componentes de variancia para cada fuente de variación del diseño. Puesto que dichas estimaciones están basadas en las técnicas tradicionales del AVAR, no explicamos aquí el procedimiento de cálculo. En la Tabla 2 se encuentran las fuentes de variación que corresponden a cada uno de los 3 ejemplos con su respectivo diseño. Hay disponibles diversos programas de ordenador que calculan los componentes de variancia, como por ejemplo el BMDP8V de Biomedical Computer Programs (Dixon & Brown, 1979), GENOVA (Brennan, 1983) y ETUDGEN (Duquesne, 1986). En todos ellos, los algoritmos y las fórmulas que presentan se refieren a planes completos; en el caso de que se disponga de planes de observación no-ortogonales que no comporten un mismo número de observaciones en los diferentes niveles de una misma faceta, Shavelson & Webb (1981) recomiendan la utilización del programa VARCOMP de Statistical Analysis System (SAS Institute, Inc. 1982), ya que consideran los planes no-ortogonales como el «talón de Aquiles» de la teoría de la generalizabilidad.

A continuación se presentan los componentes de variancia y su contribución en porcentajes correspondientes a los tres ejemplos:

Diseño 1			Diseño 2			Diseño 3	
Fuente de variación $\alpha$	Estimación del componente de variancia $\sigma^2_\alpha$	%	Fuente de variación $\alpha$	Estimación del componente de variancia $\sigma^2_\alpha$	%	Estimación del componente de variancia MIXTO $\sigma^2_{\alpha/M}$ $\sigma^2_{\alpha/M} \rightarrow E^2(\epsilon)$	%
Observadores (O)	0,0153	0,25	Fotografías (F)	-0.0501	0	-0.0542	0
Individuos (I)	0,0948	1,56	Escalas (E)	0.0300	0.45	-0.2850	0
Categorías (C)	4,5549	74,72	FE	-0.0672	0	0.0466	0.71
OI	0,1649	2,71	Tomas (T)	-0.1364	0	-0.0682	0
OC	0,0312	0,51	FT	-0.0075	0	0.0037	0
IO	0,7733	12,69	ET	-0.6300	0	-0.3150	0
IOC	0,4615	7,57	FET	0.2276	3.45	0.1138	1.74
			JT	1.4154	21.32	1.4154	21.64
			FI:T	0.1735	2.61	0.1735	2.65
			EJT	3.2850	49.48	3.2850	50.22
			FEJT	1.5074	22.71	1.5074	23.04

El análisis de los componentes nos aporta información sobre el diseño de medida. De hecho, el objetivo de un análisis de generalizabilidad se centra más bien en el análisis de los componentes que en los coeficientes de generalizabilidad. El diseño 1 nos ofrece un componente I relativamente pequeño, lo que no es muy deseable porque los individuos constituyen en este diseño la faceta de diferenciación. En cuanto a las facetas de generalización, los componentes de observadores e interacciones con los observadores son pequeños, mientras que categorías e interacciones con categorías son relativamente grandes. Este hecho sugiere que en estudios posteriores podemos reducir el número de observadores, sin que por ello reduzcamos la precisión en la generalización. Por el contrario, las categorías contribuyen al error de una manera significativa, por lo que necesitaremos aumentar el número de las mismas si deseamos un nivel alto de generalizabilidad. La modificación y optimización de este diseño será considerada de una forma más extensiva en la discusión de la tercera fase.

La estimación de los componentes del diseño 2 y 3 nos ofrece algunos valores negativos y aún cuando algebraicamente es imposible encontrar una suma de cuadrados negativa y sabiendo que una variancia es necesariamente positiva o nula, un componente de variancia puede dar lugar a una estimación negativa. Ello es debido a las fluctuaciones muestrales en la estimación de los cuadrados medios. En estos casos Cronbach et al. (1972) proponen reemplazar el valor negativo encontrado por un valor nulo, pero Cardinet et al. (1981) sugieren aceptar la estimación negativa del componente tan solo en el caso de que tengan que estimarse componentes mixtos y si éstos son negativos entonces se les podrá reemplazar por un valor nulo. Tanto en el diseño 2 como en el 3 sólo observamos valores altos de los componentes en las facetas Jueces anidados en Tomas (J:T) y en todas las interacciones en las que intervienen, lo que es bastante deseable ya que ambas facetas constituyen la diferenciación en el diseño 2 y la faceta Tomas la diferenciación en el diseño 3. Los componentes de las facetas de generaliza-

ción (fotografías y escalas) así como sus interacciones ofrecen valores bajos o nulos (dado el signo negativo), lo que sugiere la reducción de costos en el número de fotografías y escalas en próximos diseños, ya que con un número inferior podemos conseguir valores parecidos a la hora de generalizar. Tan solo las escalas (E) en interacción con las facetas de diferenciación (J:T) contribuyen al error de una manera significativa. Este hecho tan solo se produce por la heterogeneidad que supone la interacción EJ:T y lo cierto es que si no se produjera la anidación (lo que supone reducir los costos por elemento o nivel) el valor de esta triple interacción es nulo. Posteriormente comentaremos las modificaciones necesarias en ambos diseños para lograr su optimización.

El análisis de generalizabilidad propiamente dicho comienza en la fase 2 especificando uno o más diseños de medida (paso 5). Cuando se lleva a cabo un estudio G y se estiman los componentes de variancia no es necesario que el investigador especifique si las facetas son de diferenciación o de generalización o incluso si son fijas o aleatorias. Por tanto, algunos diseños se analizan con los datos del estudio G. Imaginemos, por ejemplo, que se desea conocer el impacto de la faceta observadores del diseño 1 como faceta fija. La única modificación necesaria es ajustar los componentes de variancia en el paso 4. El análisis de los ejemplos 2 y 3 podría llevarse a cabo con los datos de un sencillo estudio G (que no incluyera ni anidaciones ni facetas fijas).

En el paso 6 habría que verificar que ninguna faceta de diferenciación esté anidada en una faceta de generalización. En todo caso, se admite que toda faceta que anide a una faceta de diferenciación forme parte de la Diferenciación. De todas formas, el anidamiento de una faceta de diferenciación en una de generalización es de por sí difícilmente concebible, ya que el propósito de una faceta fija es que la suma de los efectos que ella provoca sea nula. En el diseño 3 una faceta de generalización está anidada en una faceta de diferenciación, lo que sí es posible.

La siguiente etapa (paso 7) incluye dos operaciones que están estrechamente ligadas a la presencia de facetas fijas en el plan de medida. La primera consiste en eliminar todos los componentes que incluyen una o más facetas fijas de *generalización* en sus subíndices primarios. Lógicamente si fijamos una faceta de generalización, la suma de todos los niveles de una faceta fija es nula, así como la suma de interacciones de un nivel de diferenciación con todos los niveles de una fija de generalización también es nula.

De esta forma, ningún nivel de las facetas de diferenciación queda modificado por la presencia o ausencia de una faceta de generalización fija. Los componentes que quedan, una vez eliminados los de facetas fijas de generalización, constituirán la «variancia activa» y entrarán a formar parte en las fórmulas de los parámetros de generalizabilidad. Ninguno de los 3 ejemplos que presentamos contiene una faceta fija de generalización. En definitiva, fijar una faceta de generalización reduce la variancia del error, pero en consecuencia se paga un precio muy alto, ya que se restringe el universo de generalización.

En lo que respecta a la segunda operación, es decir facetas fijas que pertenecen a la faceta de *diferenciación*, o también facetas aleatorias finitas ( $N < \infty$ ), deberán ser ajustados, para evitar un sesgo en la estimación posterior de los parámetros de generalizabilidad, todos los componentes que incluyen una faceta

fija o aleatoria finita, en su subíndice primario. Cada uno de los componentes se multiplica por un coeficiente igual a  $[N_{(f)}-1]/N_{(f)}$ , en donde  $N_{(f)}$  es el número de niveles admisibles de la faceta cuyo número es finito. Si un componente contiene muchos subíndices de facetas finitas en su subíndice primario, hará falta aplicar tantos coeficientes de ajuste como facetas existan de esa clase. Designaremos los componentes ajustados de  $\alpha$  por el término «esperanza de variancia de  $\alpha$ », que simbolizamos por  $E^2_{(\alpha)}$ . Este aspecto no aparece en la obra de Cronbach et al. (1972), dado que no tratan el problema de las facetas finitas en la diferenciación.

En nuestro diseño 3, las tomas (T) es una faceta fija de diferenciación y por tanto los componentes de variancia de las fuentes T, FT, ET y FET deberán multiplicarse por  $[N_t-1]/N_t$ , donde  $N_t$  es el número de niveles de la faceta tomas. Así  $\sigma^2_{(fet/M)}[N_t-1]/N_t = E^2_{(fet)}$ .

En el paso 8 se calcula la variancia de diferenciación (variancia de la puntuación universo). Dicha variancia es la suma de todos los componentes de la «variancia activa» que *sólo* incluyen facetas de diferenciación en su subíndice primario. En la Tabla 2 se muestran los valores de todos los componentes que contribuyen a las variancias de la puntuación universo y de los errores absoluto y relativo de los tres ejemplos. Téngase en cuenta que los componentes de variancia de diferenciación del error absoluto y relativo son diferentes en los diseños 2 y 3 aún cuando utilizamos las mismas escalas en ambos diseños. El hecho de fijar la faceta de diferenciación en el diseño 3 hace que la diferenciación sea nula y que por tanto tendremos poca precisión en la generalización.

Posteriormente se calcula la variancia de error absoluto (paso 9). Se suman todos los componentes restantes, después de haber dividido cada uno de ellos por el producto del número de niveles de facetas de generalización que aparecen en su subíndice *total*. Si aparecen facetas de generalización aleatorias finitas ( $\infty > N > n$ ) en el subíndice *primario* de un componente, se multiplicará el término obtenido anteriormente por el índice corrector del universo finito correspondiente:  $[N_{(f)}-n_{(f)}]/[N_{(f)}-1]$ . Hay que hacer notar, sobre todo cuando se analiza un componente de variancia, en un estudio G, que el valor de los componentes de error que forman parte del término de error reducirán su valor al dividirlo por el producto de niveles de las facetas de generalización. Imaginemos que el componente de la variancia de las fotografías es muy elevado utilizando 50 fotografías, pero hemos de pensar que la cantidad que aporta el término de error es el valor del componente dividido por 50.

La variancia de error relativo (paso 10) está formada por la variabilidad debida a las interacciones de las facetas de generalización aleatorias con las facetas de diferenciación. A destacar que los efectos principales (no-anidados) de las facetas de generalización no aportan nada a la variancia de error relativo. La razón estriba en que si se realizan interpretaciones relativas, los efectos de estas fuentes son constantes para cada individuo o cualquier otra faceta de diferenciación. Sin embargo, la interacción de estas fuentes con facetas de diferenciación sí que contribuyen al error relativo. Para calcularlo, es suficiente con sumar todos los componentes de variancia del error absoluto (incluso con sus coeficientes correctores) que tienen *al menos* una faceta de diferenciación en su subíndice *total*.

Finalmente se calculan los coeficientes de generalizabilidad (paso 11) dividiendo la variancia de diferenciación por la suma de la variancia de diferenciación y de la variancia de error (relativa o absoluta, según la interpretación que se desee realizar). Los coeficientes de generalizabilidad para los tres diseños son:

$$\begin{aligned} \text{Diseño 1} \quad & \begin{cases} E\sigma^2_{A(i)} = \hat{\sigma}^2_{(i)} / [\hat{\sigma}^2_{(i)} + \hat{\sigma}^2_{\Delta}] = 0.0948 / [0.0948 + 1.8449] = 0.0489 \\ E\sigma^2_{\delta(i)} = \hat{\sigma}^2_{(i)} / [\hat{\sigma}^2_{(i)} + \hat{\sigma}^2_{\delta}] = 0.0948 / [0.0948 + 0.3215] = 0.2277 \end{cases} \\ \text{Diseño 2} \quad & \begin{cases} E\sigma^2_{A(i,t)} = [\hat{\sigma}^2_{(i)} + \hat{\sigma}^2_{(i,t)}] / [\hat{\sigma}^2_{(i)} + \hat{\sigma}^2_{(i,t)}] + \hat{\sigma}^2_{\Delta} = 0.6035 \\ E\sigma^2_{\delta(i,t)} = [\hat{\sigma}^2_{(i)} + \hat{\sigma}^2_{(i,t)}] / [\hat{\sigma}^2_{(i)} + \hat{\sigma}^2_{(i,t)}] + \hat{\sigma}^2_{\delta} = 0.6054 \end{cases} \\ \text{Diseño 3} \quad & \begin{cases} E\sigma^2_{A(t)} = \hat{\sigma}^2_{(t)} / [\hat{\sigma}^2_{(t)} + \hat{\sigma}^2_{\Delta}] = 0 \\ E\sigma^2_{\delta(t)} = \hat{\sigma}^2_{(t)} / [\hat{\sigma}^2_{(t)} + \hat{\sigma}^2_{\delta}] = 0 \end{cases} \end{aligned}$$

Estos coeficientes estiman la generalizabilidad a través de las facetas de generalización (observadores, categorías) y por ejemplo en el diseño 1 si el coeficiente de generalizabilidad tuviera un valor próximo a la unidad (cosa que aquí no ocurre) podría interpretarse como la correlación entre el conjunto de 100 puntuaciones y otro conjunto de otras 100 puntuaciones obtenidas en otro conjunto de 5 observadores sobre otras 3 categorías. La magnitud del coeficiente (que varía de 0 a 1) se interpreta de la misma forma que los coeficientes de fiabilidad tradicionales, considerando como niveles «aceptables» los mismos que en los coeficientes de fiabilidad. El coeficiente es aplicado sobre la puntuación media de observadores y categorías (o sobre la suma de valores de observadores y categorías). A destacar, por tanto, que los coeficientes de generalizabilidad se refieren al universo de generalización (observadores y categorías en nuestro primer ejemplo). La misma interpretación sirve para los otros dos ejemplos, el diseño 2 con dos facetas de generalización y el 3 con tres facetas de generalización.

Los mayores beneficios de un análisis de generalización se derivan esencialmente de la tercera fase cuando se llevan a cabo modificaciones en el diseño de medida y se elige un diseño optimizado (óptimo en el sentido de que se busca una máxima generalizabilidad dentro de los costos u otras restricciones prácticas o, alternativamente, que se reduzcan los costos mientras se mantenga un elevado o aceptable nivel de generalizabilidad). A continuación presentamos diferentes ilustraciones de los tipos de modificación que pueden realizarse en nuestros tres ejemplos.

### *Análisis de generalizabilidad: optimización*

El coeficiente de generalizabilidad relativo del diseño 1 (0,2277) es muy pequeño, lo que indica poca precisión de generalización de observadores y categorías cuando el objeto de medida (puntuación universo) son los individuos. Evidentemente, si no se produce un cambio en el diseño de medida, ello implicará una serie de modificaciones en el plan de estimación. Si por el contrario, se hubiera obtenido un valor cercano a la unidad se podrían reducir los costos de la

investigación, por ejemplo, reduciendo el número de observadores o de categorías. De hecho, el número de observadores nunca suele ser más de dos. Si se reduce el número de observadores a 2 la estimación implicaría cambiar el valor de  $n_o = 5$  a  $n_o = 2$  en la fórmula de la variancia de error relativo (o de error absoluto si consideráramos el coeficiente de generalizabilidad absoluto). Probablemente ello implicaría una reducción del coeficiente de generalizabilidad, pero aún así sería aceptable en la mayoría de estudios D.

La Tabla 4 presenta el diseño 1 con el plan de medida original, pero modificando sucesivamente el plan de estimación para lograr una optimización de cada una de las facetas en combinación con las otras facetas y así obtener una precisión en la generalización adecuada a este tipo de investigaciones. Se verifica evidentemente que las medidas relativas son más generalizables que las medidas absolutas; de hecho su ambición es menos pretenciosa. En cuanto a la significación de estos resultados, podemos comprobar que las generalizabilidades no son muy buenas en este plan de medida en que los individuos (I) constituyen el objeto de estudio. El hecho de limitar la población de observadores, fijando la faceta de observadores, no consigue aumentar la precisión de generalización (0.69) y tan solo conseguiríamos una buena precisión aumentando a 100 observadores (0.91), lo que implicaría que las otras dos facetas también sean aleatorias finitas ( $N_i = N_c = 100$ ), pero en tal caso los costos de esta modificación serían altísimos.

TABLA 4. OPTIMIZACIÓN DEL DISEÑO 1

Faceta	Niveles observados	Niveles estimados	Modificaciones al plan de estimación														
O	$n_o = 5$	$N_o = \infty$	5	10	50	100	100	100	100	100	100	100	100	100	5	5	20
I	$n_i = 4$	$N_i = \infty$	100	100	100	100	5	50	100	100	100	100	100	4	4	4	4
C	$n_c = 3$	$N_c = \infty$	100	100	100	100	100	100	1	5	50	200	0	0	0	0	3
	$E\sigma_d^2 = 0.2277$		0,69	0,79	0,89	0,91	0,91	0,91	0,91	0,99	0,38	0,85	0,72	0,60	0,26		
	$E\sigma_A^2 = 0,0489$		0,51	0,57	0,62	0,63	0,63	0,63	0,99	0,08	0,47	0,60	0,30	0,05			

Si aumentamos el número de individuos de 5 a 50 tampoco conseguiremos una mejora de la precisión (0.91), pero es lógico ya que los individuos constituyen la diferenciación y no es a ellos a quien tenemos que generalizar. Es evidente que si tuviéramos una sola categoría para evaluar a los individuos no se produciría heterogeneidad en los mismos y por tanto la precisión sería muy alta (0.99), pero en consecuencia hemos restringido el universo de generalización a una única categoría. Si aumentamos el número de categorías a 50 el coeficiente sería relativamente alto (0.85), pero en consecuencia aumentarían los costos de la investigación. Si fijamos las facetas observadores e individuos tampoco conseguimos precisión suficiente (0.72 y 0.60), al igual que si fijamos individuos y categorías (0.26). En consecuencia, la heterogeneidad de las categorías (puesta ya de mani-

fiesto en el elevado componente de variancia y en su alta contribución al error, 4,55 y 74,72 respectivamente) no permite una generalización precisa a las mismas y por tanto las modificaciones al plan de estimación no son suficientes y se necesitaría una revisión de las mismas para plantear con éxito investigaciones futuras.

En este mismo diseño podríamos formular otros planes de medida diferentes, como por ejemplo el plan (C/-/-/I,O) donde las categorías constituirían la faceta de diferenciación, mientras que individuos y observadores las facetas de generalización. Probablemente el coeficiente de generalizabilidad sería alto ya que la estimación de  $\sigma^2_c$  a través de las tres categorías sobrevaloraría la heterogeneidad de éstas y en consecuencia crecería artificialmente la variancia de diferenciación. El plan (I,C/-/-/0) nos permitiría verificar la fiabilidad interobservadores. El plan (O/-/-/IC) permitiría diferenciar a los observadores, es decir el grado de evaluación de  $n_o$  observadores cualesquiera a partir de los valores que atribuyen a una muestra aleatoria de  $n_i$  individuos, a los que registran a través de  $n_c$  categorías aleatorias. El plan (I,O/-/-/C) evaluaría la precisión de las medidas (absolutas o relativas) de las medias que han dado los observadores a los sujetos evaluados en  $n_c$  categorías aleatorias. En definitiva, son muchos planes de medida diferentes los que podemos realizar y sobre los que podemos realizar modificaciones al plan de estimación y optimizaciones al plan de medida. La elección de uno u otro dependerá en gran medida de los objetivos de la investigación, objetivos que en algunos casos limitarán el número de planes de medida. Lo que sí debe quedar claro es que en estudios observacionales del comportamiento podremos determinar la fiabilidad de la precisión de las medidas, la fiabilidad inter e intraobservadores, la validez de las categorías, la precisión del tiempo de registro, la evaluación de los momentos diferentes, etc. y todo ello de forma conjunta en un único diseño que nos permite un coeficiente de generalizabilidad diferente para cada tipo de interpretación. El conocimiento de la generalizabilidad de un dispositivo permite naturalmente apreciar mejor los resultados: podemos saber en qué medida nos podemos fiar de los valores obtenidos; y podemos ver en qué dirección podemos mejorar nuestras investigaciones.

Consideremos ahora nuestro segundo ejemplo, el plan de observación (J:T)  $\times$  F  $\times$  E. La Tabla 5 presenta diferentes planes de optimización al diseño original que proponemos en la Tabla 2. En cuanto a la significación de estos resultados podemos comprobar que si fijamos las facetas T y J:T (que son la diferenciación) necesitamos aumentar el número de niveles de 10 a 100 para conseguir una precisión muy alta en las facetas de generalización.

Concretamente con  $N=10$  el coeficiente relativo es 0,80, con  $N=50$  es 0,95 y con  $N=100$  es 0,98. La optimización es excelente pero el precio que se paga es muy alto, ya que tenemos que aumentar el número de observaciones de 240 a 100.000 y por tanto el costo de la investigación no corresponde a una buena optimización. Si por el contrario fijamos las facetas de generalización el coeficiente es bajo (0,60) y además restringimos el universo de generalización. Comprobamos también que manteniendo las facetas aleatorias finitas ( $N=100$ ) obtenemos excelentes resultados, pero siempre con costos muy altos. El último plan optimizado supone aumentar todas las facetas a  $N=10$ , lo que supone no incrementar en exceso los costos y obtener una buena precisión absoluta y relativa (0.79

TABLA 5. OPTIMIZACIÓN DEL DISEÑO 2

Faceta	Niveles observados	Niveles estimados	Modificaciones al plan de estimación									
			10	50	100	6	100	100	100	100	100	10
F	$n_f = 6$	$N_f = \infty$	10	50	100	6	100	100	100	100	10	
E	$n_e = 4$	$N_e = \infty$	10	50	100	4	100	100	100	100	10	
T	$n_t = 2$	$N_t = \infty$	2	2	2	10	100	100	4	10	10	
J/T	$n_j = 5$	$N_j = \infty$	5	5	5	10	10	50	10	10	10	
$E\phi_{\delta}^2 = 0,605$			0,80	0,95	0,98	0,60	0,98	0,98	0,98	0,98	0,80	
$E\phi_{\Delta}^2 = 0,603$			0,79	0,95	0,98	0,60	0,98	0,98	0,98	0,98	0,79	

y 0.80 respectivamente) que en muchos estudios D puede ser considerada como suficiente aunque no deseable. En definitiva, las anidaciones que se presuponen en este ejemplo para reducir los costos en el número de niveles de las facetas no determinan una buena precisión y por el contrario nos hacen aumentar el número de niveles de las facetas que nosotros en principio habíamos restringido con la anidación.

En cuanto al diseño 3 es evidente que al fijar la faceta de diferenciación (con un valor muy pequeño,  $n=2$ ) carecemos totalmente de precisión y es obvio que no da lugar ni a modificaciones ni a optimizaciones. En todo caso se ha de replantear de nuevo la investigación aumentando como mínimo el número de niveles de la faceta de diferenciación (T). El análisis de generalizabilidad ha servido de estudio piloto para plantear de nuevo una recogida de datos más coherente al diseño de medida.

En nuestro ejemplo 2 también podría plantearse el estudio de otros diseños de medida para verificar cuáles son las facetas que nos permiten mayor generalización. A título de ejemplo hemos seleccionado dos planes de medida que no lleven consigo facetas anidadas. El primero de ellos, el plan (-/J,E/F/T), en el que solamente la faceta T es aleatoria infinita consigue una excelente generalización de fotografías y tomas (0.99 tanto para el coeficiente absoluto como para el relativo). Los resultados son en cierto modo lógicos ya que se ha restringido el universo de las fotografías al número de niveles observados. Ello nos sugiere que el número de niveles de las tomas (T) ha de ser aumentado, ya que en el ejemplo que proponemos es la única faceta infinita y además de generalización.

El segundo plan (E/J-/T,F), con  $N_e=100$ ,  $n_j=5$ ,  $N_t=\infty$ ,  $N_f=100$  nos ofrece también la base de un excelente estudio D con generalizabilidad relativa de 0.94 y absoluta de 0.92. Es decir nos permite tener un universo más amplio de generalización incluso sabiendo que el número de jueces (J) es fijo en la diferenciación.

Por tanto, el plan de medida óptimo de un estudio D no será casi nunca el del estudio G correspondiente. Ante todo un estudio de generalizabilidad tiene como primera misión estimar la importancia de todas las fuentes de variancia que afectan a los datos y no la de vigilar la diferenciación máxima de ciertos objetos de estudio. El número de situaciones de decisiones posibles, a partir de las

facetas descritas por un mismo estudio G, es muy grande y hará falta adaptar los planes de medida a cada una de las situaciones.

En el caso de coeficientes de generalizabilidad con valores bajos e inaceptables, una solución bastante común y muy utilizada para conseguir que el plan aumente la generalizabilidad en los subsiguientes estudios D es la de incrementar el número de niveles de ciertas facetas. Los valores de los componentes de variancia y la fórmula de la variancia del error nos indicarán qué facetas contribuyen sustancialmente al error. Al incrementar los niveles de estas facetas se producirá un importante aumento de la generalizabilidad.

Estos ejemplos han servido para ilustrar una de las grandes ventajas de un análisis de generalizabilidad: la capacidad de diseñar estudios D más eficientemente en base a la información aportada por un estudio G. Si tenemos en cuenta la relación nivel de precisión-costos es obvio que podremos diseñar un estudio D *óptimo*. La utilización de un análisis de generalizabilidad para modificar diseños sucesivos a largo plazo ha sido estudiada en el trabajo de Johnson & Bell (1985).

Cuando se realizan estudios G es necesario difundir los valores de los componentes de variancia para que otros usuarios puedan diseñar su propio estudio D. También deberá ser especificado claramente el diseño de medida y el universo de generalización con el fin de evitar la ambigüedad del coeficiente de generalizabilidad resultante.

## CONCLUSIÓN

Cronbach et al. (1972) han desarrollado un modelo que admite la multidimensionalidad de las fuentes de variancia que afectan a una puntuación observada. La medida observada representa una muestra aleatoria de un conjunto de medidas que varían en función de los observadores, categorías, momentos, ocasiones, etc. Puesto que no comporta otro tipo de indeterminación que el carácter aleatorio de elección de niveles observados, la medida se presta por tanto a los tratamientos estadísticos habituales.

Una forma de organizar los diferentes tipos de aplicación de esta teoría sería por ejemplo oponer las aplicaciones «a priori» a las «a posteriori»:

— En el espíritu de Cronbach et al. (1972), un análisis de generalizabilidad constituye normalmente un estudio piloto, que sirve para preparar una experiencia a una más grande escala. El trabajo previo de estimación de las fuentes de variancia debe permitir poner a punto los dispositivos de medida adaptados a las decisiones consideradas en la investigación principal (plan de optimización). De todas formas, todas las fases de la generalización constituyen de por sí una puesta a punto: redefinición del universo de generalización, purificación de la diferenciación, fijación de las facetas que inducen a un sesgo excesivo, etc.

— El hecho de que un análisis haya sido hecho «a posteriori» no significa que éste no tenga influencias en las investigaciones posteriores. Por el contrario, los investigadores retoman los conceptos y los instrumentos que se han revelado útiles en los trabajos de sus predecesores para así conseguir una mejora progresiva de sus dispositivos de observación. Los análisis de generalizabilidad permiten

racionalizar estos procesos de selección y desarrollar los instrumentos de medida.

Finalmente, y sin ningún tipo de duda por el tipo de modelo conceptual que aporta, esta teoría marcará el desarrollo futuro de las investigaciones científicas en las diversas áreas de las ciencias del comportamiento. No creemos que sea posible plantear un procedimiento de observación y pretender obtener una serie de valores si antes no nos hemos hecho las siguientes preguntas: ¿cuál es el universo de condiciones de observación al que pretendemos generalizar esos valores? y ¿cuál es la población de objetos de estudio a la que este procedimiento de observación parece ser aplicable?

## RESUMEN

Vamos a tratar de sintetizar la contribución original de los autores de la primera obra sobre generalizabilidad (Cronbach, Gleser, Nanda y Rajaratnam, 1972) y de cómo han ido progresando los conceptos anteriores de la fiabilidad de las medidas dentro del contexto particular de la metodología observacional. El primer objeto de estos autores ha sido poner en orden la multitud de coeficientes que se habían desarrollado para cuantificar la fiabilidad de un instrumento de medida. Sin enumerarlos, recordemos simplemente que estos coeficientes tienen su fundamento en una comparación entre observadores, individuos, momentos, etc. y que buscan, desde grados muy diversos, una hipótesis de equivalencia. Estos coeficientes parecen entrar en contradicción entre ellos y no son capaces de medir la fiabilidad real de un instrumento. Cronbach y cols. han desarrollado un modelo más extenso que admite la multidimensionalidad de las fuentes de variancia que afectan a una puntuación observada.

## SUMMARY

We are trying to summarize the original contribution of those authors who, for the first time, spoke of generalizability (Cronbach, Gleser, Nanda & Rajaratnam, 1972) and also how the early concepts of reliability of the measures have developed in the specific context of the observational methodology. The first aim of these authors has been to put order into the multitude of coefficients which have been developed gradually to quantify the reliability of a measurement instrument. Without naming them, simply let's remember that these coefficients have their grounds in a comparison between observers, people, occasions, etc. and therefore, from very different angles, they are seeking an equivalent hypothesis. These coefficients seem to contradict each other and they cannot measure the real reliability of an instrument. Cronbach *and cols.* have developed a wider model which allows for the multidimensionality of the variance sources which affect the observed scores.

## REFERENCIAS

- Arnau Gras, J. (1978). *Psicología Experimental. Un enfoque metodológico*. México: Trillas.
- Arnau Gras, J. (1981). *Diseños experimentales en psicología y educación*. Vol. 1. México: Trillas.
- Anguera Argilaga, M.T. (1983). *Manual de prácticas de observación*. México: Trillas.
- Anguera Argilaga, M.T. (1988). *Observación en el aula*. Barcelona: Graó.
- Anguera Argilaga, M.T. y Blanco Villaseñor, A. (1984, septiembre). *Aplicación de la teoría de la generalizabilidad a datos observacionales*. Comunicación presentada al XXIII Congreso Internacional de Psicología. Acapulco, México.
- Berk, R.A. (1979). Generalizability of behavioral observations: A clarification of interobserver agreement and interobserver reliability. *American Journal of Mental Deficiency, 83*, 460-472.
- Blanco Villaseñor, A. (1983). *Análisis cuantitativo de la conducta en sus contextos naturales*. Tesis Doctoral no publicada, Universidad de Barcelona.
- Blanco Villaseñor, A. (1986a). *Problems of generalizability in Environmental Psychology*. Paper presented at the 21st International Congress of Applied Psychology. Jerusalem, Israel.
- Blanco Villaseñor, A. (1986b). *Generalizabilidad en diseños de observación de la conducta*. Comunicación presentada en la 1ª Jornada de Psicología de la Delegación Catalana de la Sociedad Española de Psicología.
- Blanco Villaseñor, A. (1986c). *Generalizabilidad de la observación de la conducta*. Trabajo inédito no publicado. Barcelona: Universidad de Barcelona. Departamento de Psicología Experimental.
- Blanco Villaseñor, A. y Anguera Argilaga, M.T. (1984, septiembre). *Fiabilidad, precisión y validez de los registros observacionales*. Comunicación presentada al XXIII Congreso Internacional de Psicología. Acapulco, México.
- Brennan, R.L. (1980). Applications of Generalizability Theory. In R.A. Berk (Ed.). *Criterion-referenced Measurement: The state of the art*. Baltimore, Md.: The Johns Hopkins University Press.
- Brennan, R.L. (1983). *Elements of generalizability theory*. Iowa City, Ia.: The American College Testing Program.
- Cardinet, J. & Tourneur, Y. (1985). *Assurer la mesure*. Berne: Peter Lang.
- Cardinet, J., Tourneur, Y. & Allal, L. (1976). The symmetry of generalizability theory: Applications to educational measurement. *Journal of Educational Measurement, 13* (2), 119-135.
- Cardinet, J., Tourneur, Y. & Allal, L. (1981). Extension of generalizability theory and its applications in educational measurement. *Journal of Educational Measurement, 18* (4), 183-204.
- Cronbach, L.J., Rajaratnam, N. & Gleser, G.C. (1963). Theory of generalizability: a liberalization of reliability theory. *British Journal of Mathematical and Statistical Psychology, 16*, 137-163.
- Cronbach, L.J., Gleser, G.C., Nanda, H. & Rajaratnam, N. (1972). *The dependability of behavioral measurements: theory of generalizability for scores and profiles*. New York: Wiley.
- Dixon, W.J. & Brown, M.B. (Eds.) (1979). *BMDP-79 Biomedical Computer Programs P Series*. Los Angeles, Ca.: University of California Press.
- Duquesne, F. (1986). Développement sur micro-ordinateur d'un programme pour l'étude de la généralisabilité des données. *Scientia Paedagogica Experimentalis, 23* (1), 29-36.
- Johnson, S. & Bell, J.F. (1985). Evaluating and predicting survey efficiency using generalizability theory. *Journal of Educational Measurement, 22*, 107-119.
- Kane, M.T. & Brennan, R.L. (1980). Agreement coefficients as indices of dependability for domain-referenced tests. *Applied Psychological Measurement, 4* (1), 105-126.
- López Feal, R. (1986). *Construcción de instrumentos de medida en Ciencias Conductuales y Sociales*. Barcelona: Alamex.
- Marcoulides, G.A. (1989). The application of generalizability analysis to observational studies. *Quality & Quantity, 23*, 115-127.
- Martínez Arias, M.R. (1981). Principios psicométricos de las técnicas en evaluación conductual. En R. Fernández Ballesteros, y J.A.I. Carrobes (Dir.), *Evaluación Conductual. Metodología y Aplicaciones* (pp. 157-198). Madrid: Pirámide.
- Medley, D.M. & Mitzel, H.E. (1963). Measuring Classroom Behavior by Systematic Observation. In N.L. Gage (Ed.). *Handbook of Research on Teaching* (pp. 247-328). Chicago, Ill.: Rand McNally.
- Mitchell, S.K. (1979). Interobserver Agreement, Reliability and Generalizability of Data Collected in Observational Studies. *Psychological Bulletin, 86* (2), 376-390.
- Nussbaum, A. (1984). Multivariate Generalizability Theory in Educational Measurement: An empirical Study. *Applied Psychological Measurement, 8* (2), 219-320.
- Rentz, J.O. (1987). Generalizability Theory: A Comprehensive method for assessing and improving the dependability of marketing measures. *Journal of Marketing Research, 24*, 19-28.

- Rowley, G.L. (1976). The reliability of observational measures. *American Educational Research Journal*, 13, 51-59.
- SAS Institute, Inc. (1982). *SAS User's Guide: Statistics*. Cary, N.C.: SAS Institute, Inc.
- Shavelson, R.J. & Webb, N.M. (1981). Generalizability theory: 1973-1980. *British Journal of Mathematical and Statistical Psychology*, 34, 133-166.
- Shavelson, R.J., Webb, N.M. & Rowley, G.L. (1989). Generalizability Theory. *American Psychologist*, 44 (6), 922-932.
- Smith, P.L. & Teeter, P.A. (March, 1982). *The Use of Generalizability Theory with Behavioral Observation*. Paper presented at the Annual Meeting of the American Educational Research Association. New York.
- Suen, H.K., Lee, P.S.C. & Owen, S.V. (in press). The effects of autocorrelation on single-facet crossed-design generalizability assessment. *Psychological Bulletin*.