Revista Mexicana de Análisis de la Conducta
Núm. Monográfico, Vol. 17, No. 3, 1991, pp. 23-63

La teoría de la generalizabilidad aplicada a diseños observacionales

Generalizability theory applied to observational designs

Angel Blanco-Villaseñor
Universidad de Barcelona.

RESUMEN

Este estudio empírico de aplicación de la teoría de la generalizabilidad viene a ilustrar la ampliación que con respecto a un diseño general de investigación, implica el principio de *simetría de las facetas o variables* de un Plan de Observación.

Cronbach, Gleser, Nanda, & Rajaratnam (1972), han desarrollado *la teoría de la generalizabilidad*, asumiendo que hay otras fuentes de variación además de las diferencias individuales e integrando cada una de estas fuentes de variación en una estructura global, que permite aplicaciones particulares de la teoría estadística del muestreo. La teoría de la generalizabilidad reconoce explícitamente las múltiples fuentes de error de medida en un diseño de investigación observacional (individuos, observadores, categorías, sesiones,...). Podemos estimar cada una de estas fuentes de error así como las diferentes interacciones entre ellas. El error de medida no es más que el efecto de las fluctuaciones debidas a la elección aleatoria de los individuos, observadores, categorías, sesiones..., es decir al muestreo de niveles particulares en cada una de las facetas (variables) del universo de observaciones posibles. Optimizar dicha medida es adaptar nuestro diseño para reducir al máximo la variancia del muestreo debida a estas facetas.

En el estudio empírico que presentamos (8 individuos han sido evaluados por dos observadores, en una situación interactiva de juego en el patio de un colegio, en un sistema de siete categorías a través de cinco sesiones de observación diferentes realizadas en diversos días) y en otras situaciones de medida frecuentemente no se trata de diferenciar individuos, sino más bien de diferenciar observadores, categorías, sesiones, etc. Ello significa asumir el principio de *simetría*, es decir que sucesivos objetos de medida pueden ser evaluados dentro de un mismo diseño. Mediante dicho principio, cada faceta (variable) de un diseño de investigación puede ser seleccionada 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. Esta sería la *diferencia* a una solución experimental que tan sólo diferenciaría a tratamientos o condiciones.

Palabras clave: Teoría de la generalizabilidad, metodología observacional.

Generalizability theory applied to observational designs

ABSTRACT

The application of generalizability theory to the general research designs is reviewed. The theory assumes that different sources of variance in a global structure permits particular applications of sampling statistical theory. Generalizability theory assumes multiple sources of measurement error in an observational design. It is possible to assess each of this sources of error and their diferent interactions. The measurement error is the effect of the fluctuations determined by the randomized choice of subjects, observers, categories, and sessions. To optimize such measure means to adapt one's design in order to decrease the sample variance determined by such facets. One example is presented where eight subjects were evaluated by two observers, in an interactive play situation, with a seven categories catalogue. Under some circumstances, differentiating individuals, is not as important as differentiating observers, categories and sessions, in order to assume the simetry principle: the successive objects of measurement may be evaluated within the same design. Under this principle, each facet of the research design may be selected as an object of study and in each generalizability analysis, this facet could be considerated as an measurement instrument or evaluative condition in the study of the remaining facets.

Key words: Generalizability theory, observational methodology.

INTRODUCCION [1]

El acto de la medición es uno de los componente esenciales de la investigación científica, tanto en las ciencias naturales como en las sociales, de la salud o del comportamiento. Ciertamente, la medición juega un rol importante en la investigación en ciencias del comportamiento, al igual que en otras disciplinas científicas. Sin embargo, la medición en disciplinas de laboratorio no suele presentar dificultades inherentes. De la misma forma que en otras ciencias naturales, la medición es una parte fundamental de la disciplina y una aproximación al desarrollo de instrumentos apropiados. Los juicios subjetivos juegan un rol menor en el proceso de medición; cualquier intento de replicación o validación son posibles gracias a soluciones tecnológicas. Hemos de mencionar, sin embargo, que un equipamiento costoso no es, en sí mismo, una forma de eliminar errores de medida.

La medición en las ciencias sociales, de la salud y del comportamiento tiene como punto de referencia a la teoría de los tests, en la que como veremos posteriormente están basados los diferentes coeficientes de fiabilidad. En las mismas se presupone que una puntuación observada, en una determinada prueba, puede descomponerse en una puntuación "verdadera" (que realmente no conocemos) y en una puntuación del "error". Dicha suposición lleva directamente a la formulación de un coeficiente de fiabilidad como la razón entre la variancia verdadera y la variancia (verdadera + error).

Hay diferentes formas de estimar la fiabilidad, como especificaremos a continuación, y cada una de ellas genera un coeficiente diferente. Podemos verificar las puntuaciones dadas por un mismo observador en dos momentos diferentes a la misma sesión de observación *(intraobservadores)*, o por diferentes observadores en el mismo período temporal *(interobservadores)*, o en diferentes ocasiones separadas por un corto intervalo de tiempo *(test-retest)*, o utilizar diferentes formas de una escala *(formas paralelas)* y así sucesivamente. Sin embargo, estas medidas *standard* no agotan todas las posibles fuentes de variación (Chalmers & Knight, 1985; Miller, 1991; Nussbaum, 1984; Schoroeder, 1984; Streiner & Norman, 1989). El objetivo del presente trabajo es precisamente utilizar una nueva vía de expresar esta variabilidad a través de los conceptos del análisis de la variancia.

Este original enfoque fue diseñado inicialmente por Cronbach, Gleser, Nanda, & Rajaratnam (1972) y es conocido como la *teoría de la generalizabilidad*. La esencia de la teoría es el postulado de que en cualquier situación de medida existen múltiples (de hecho infinitas) fuentes de variación (denominadas *facetas* en dicha teoría). Uno de los objetivos importantes de la medición es intentar identificar y medir los componentes de variancia que están aportando error a una si-

1 Quisiera expresar mi más profundo agradecimiento a Carlos Santoyo Velasco, Profesor de la Universidad Nacional Autónoma de México, por sus sugerencias y comentarios críticos, realizados durante su estancia en período sabático en la Universidad de Barcelona (Departamento de Metodología de las Ciencias del Comportamiento).

tuación y, entonces, implementar estrategias que reduzcan la influencia de estas fuentes de error sobre la medida. En definitiva, estamos ante un elegante y práctico enfoque para entender las diferentes fuentes de variación que pueden estar afectando a un dato observacional.

Nociones básicas de fiabilidad

El concepto de *fiabilidad* es, igualmente, muy simple. Antes de tener la seguridad de que un instrumento mide aquello que nos proponemos es necesario, en primer lugar, acumular evidencia de que la escala está midiendo *algo* de alguna manera reproducible. Es decir, la primera fase para obtener la evidencia del valor de un instrumento es demostrar que las mediciones de los individuos en diferentes ocasiones, o por diferentes observadores, o en similares sesiones, producen los mismos o similares resultados.

Esta es la idea básica que esconde dicho concepto: un índice de la dimensión por lo cual las mediciones de los individuos obtenidos en circunstancias diferentes nos ofrecen resultados similares. Sin embargo, el concepto es más refinado en la teoría clásica de la medida. Por ejemplo, consideremos la fiabildad de las escalas de las básculas de baño; es suficiente indicar que las escalas son precisas en + 1 kg. Disponiendo de esta información podemos juzgar si las escalas son adecuadas para distinguir entre adultos (probablemente sí) o para evaluar el aumento de peso de niños prematuros (probablemente no), dado que tenemos información previa del promedio y variación del peso de adultos y niños prematuros.

Dicha información no está disponible en el desarrollo de escalas subjetivas. Cada escala produce mediciones diferentes unas de otras. Por tanto, si especificamos que una determinada escala es precisa en ± 3.4 unidades no estamos indicando su valor de medición de los individuos, a menos que tengamos alguna idea sobre el rango posible de puntuaciones en la escala. Para solventar este problema, la fiabilidad es generalmente definida como la razón entre la variabilidad-individual y la variabilidad total de las puntuaciones; en otras palabras, la fiabilidad es una medida de la proporción de variabilidad en las puntuaciones que es debida a las diferencias verdaderas entre los individuos. Así, la fiabilidad viene expresada como un número entre 0 y 1, indicando el valor 0 que no existe fiabilidad y el valor 1 la fiabilidad perfecta.

Uno de los principios importantes de la fiabilidad de un instrumento es la forma de obtener los datos para el cálculo posterior de un coeficiente de fiabiliadad. En primer lugar, dado que la fiabilidad implica la razón de variabilidad entre sujetos con respecto a la variabilidad total, se hace necesario llevar a cabo un estudio en una muestra extremadamente heterogénea y asegurar que la muestra utilizada en dicho estudio sea exactamente la misma que la que deseamos estudiar. En segundo lugar, dado que disponemos de varias formas de obtener medi-

das de fiabilidad, la magnitud del coeficiente de fiabilidad será el reflejo directo del enfoque utilizado. Describimos a continuación algunas definiciones abiertas:

1.- *Consistencia interna*. Las medidas de consistencia interna están basadas en una única administración de la medida. Si la misma dispone de un número relativamente grande de items para la comprensión de una dimensión, es razonable esperar que las puntuaciones en cada ítem correlacionen con las puntuaciones de todos los otros items. Esencialmente, las medidas de consistencia interna reflejan la proporción de la correlación entre todos los items de una medida. Las diversas formas de calcular estas correlaciones son denominadas *alfa de Cronbach, Kuder-Richardson* o *método de las dos mitades*, aunque todas ofrecen resultados similares. Aunque la técnica implica una única administración, no es difícil obtener dichos coeficientes. Sin embargo, no tienen en cuenta cualquier variación de un día a otro, de un observador a otro o de una situación a otra, y por tanto no nos ofrecen una interpretación optimista de la fiabilidad verdadera.

2.- *Estabilidad*. Disponemos de diferentes formas de examinar la reproducción de una medida administrada en diferentes ocasiones. Por ejemplo, comprobar el grado de acuerdo entre diferentes observadores (*fiabilidad inter-observadores*); el acuerdo entre observaciones realizadas por el mismo observador en dos ocasiones diferentes (*fiabilidad intra-observadores*); las observaciones de un mismo individuo o en dos ocasiones separadas por un intervalo de tiempo (*fiabilidad test-retest*). Como mínimo, cualquier decisión sobre el valor de una medida estará basada en algún tipo de información acerca de la estabilidad del instrumento.

3.- *Standards de fiabilidad*. Una de las dificultades de los coeficientes de fiabilidad es que representan un número entre 0 y 1, y no ofrecen interpretaciones comunes. Algunos autores han realizado recomendaciones acerca del nivel mínimo aceptable de fiabilidad. Ciertamente, los valores de consistencia interna deben ser mayores que 0.8 y es razonable ofrecer medidas de estabilidad si exceden el valor 0.5. Dependiendo de la utilización y del costo de la interpretación es evidente que se requerirán valores más altos.

Finalmente, si suponemos que cualquier tipo de respuesta tiene asociado algún tipo de error de medida, es posible promediar o sumar dichas respuestas sobre una serie de cuestiones con el fin de reducir este error. Por ejemplo, si el instrumento original ofrece un valor de fiabilidad de 0.5, doblando el número de cuestiones aumenta la fiabilidad a 0.67 y cuadriplicando a 0.8. Como resultado, reconocemos que la brevedad no es necesariamente un atributo deseable en estos instrumentos utilizados en la investigación psicológica.

Fiabilidad de los registros observacionales

En las ciencias del comportamiento, los fenómenos observados están influidos por tal cantidad de factores que una repetición de una misma experiencia o la utilización de cualquier otro instrumento pueden modificar considerablemente

el resultado que se obtuvo la primera vez. Por ello, la actitud científica más elemental nos lleva a preguntarnos si esos valores observados son interpretables o si, por el contrario, son el resultado de fluctuaciones alcatorias, introducidas por la propia medida. Este interrogante es particularmente más necesario en los diseños de observación de la conducta (Anguera, 1983, 1987, 1988, 1991; Blanco y Anguera, 1984; Berk, 1979; Rowley, 1976; Shadish, Cook, & Leviton, 1991), dado que muchos trabajos de investigación no dejan de poner en duda el valor de los registros observacionales asociados.

Un instrumento es fiable si tiene pocos errores de medida, si 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 (tests) han ido reemplazándose poco a poco (en psicología clínica, escolar, evolutiva) por observaciones de los individuos en situaciones naturales o cuasi-naturales. Aun cuando estos estudios observacionales varían completamente en contenido y método, todos ellos utilizan observadores humanos para registrar el comportamiento de los individuos. Sorprendentemente, sin embargo, la fiabilidad de los métodos observacionales no ha recibido la misma atención que la fiabilidad de los métodos más tradicionales (Johnson & Bolstad, 1973).

Existen al menos tres formas de entender la *fiabilidad* de los datos observacionales (Blanco, 1983, 1986 b, 1989; Medley & Mitzel, 1963; Mitchell, 1979):

1.- Nos referimos 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 un mismo momento ("coefficient of observer agreement").

Estos coeficientes se interpretan como la fiabilidad del instrumento que se ha utilizado. El índice más comúnmente utilizado para valorar la calidad de estos registros observacionales es el porcentaje de acuerdo interobservadores, que, como su nombre indica, es el porcentaje de unidades (temporales en el caso de registro por intervalos) a través de las cuales los registros de dos observadores están en acuerdo con el registro de la conducta (Anguera, 1983, 1985, 1988). Se han definido otros muchos índices para este tipo de datos, basados en la información nominal u ordinal que proporcionan, y que podemos obtener a través de tablas de contingencia 2 x 2, correlacionales ordinales u otro tipo de correlaciones (como por ejemplo el clásico índice "kappa").

2.- Una medida observacional podría considerarse como un caso especial de una prueba psiscológica estandarizada, y en tal caso podemos utilizar las definiciones de fiabilidad de la teoría psicométrica clásica, 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, 1989):*

Intraobservadores: Un único observador en dos momentos diferentes, es decir diferentes observaciones de un mismo comportamiento, pero sin interrupción temporal, ya que es la misma sesión de observación, grabada mediante soportes audio y/o video, observada dos veces. El error reflejaría las inconsistencias del observador al utilizar un sistema de categorías.

Interobservadores: Dos observadores registran una misma sesión que previamente ha sido grabada mediante soportes audio y/o video. No necesariamente se exige un registro simultáneo, pues la observación está grabada y por tanto tampoco existe una interrupción temporal. No hay que confundir con la concordancia interobservadores, que sí exigiría un registro simultáneo de la observación directa, además de que su expresión es mediante un porcentaje de acuerdo y no a través de un coeficiente de correlación.

b) *Equivalencia* (formas paralelas o equivalentes de sesiones de observación) y *homogeneidad* (dos mitades o partes de una misma sesión de observación), es decir obtener puntuaciones de dos intrumentos similares o de dos partes del mismo instrumento, respectivamente. En el primer caso tendríamos dos observadores registrando dos sesiones de observación muy similares (por ejemplo, primera y segunda media hora de una clase). En el segundo caso, se trataría de dos observadores registrando, en un mismo período temporal, subdivisiones de una misma sesión de observación (minutos pares e impares de una sesión de observación).

c) *Constancia* (o estabilidad), es decir obtener puntuaciones del mismo instrumento en dos momentos diferentes, pero con una interrupción temporal. Se refiere a las observaciones realizadas por el mismo observador en momentos diferentes y expresa por tanto la estabilidad de la conducta del observador en el tiempo, pero no entendida desde el punto de vista longitudinal (por ejemplo, un observador registra en dos días diferentes la técnica de instrucción que utiliza un profesor para impartir sus clases).

Las diferencias entre *concordancia* (acuerdo) y *fiabilidad* (correlación) se basan en la forma en que se definen estos índices. Los coeficientes de fiabilidad dividen la variancia de un conjunto de puntuaciones en una puntuación verdadera (diferencias individuales) y un componente de error. Los porcentajes de acuerdo interobservadores, sin embargo, no aportan información sobre las diferencias individuales entre sujetos y tan sólo contienen información de una sola de las posibles fuentes de error (diferencias entre observadores).

Tampoco los coeficientes de fiabilidad son todo lo perfectos que podrían llegar a ser. Los coeficientes que utilizan dos puntuaciones de un mismo instrumento (fiabilidad, intraobservadores e interobservadores) confunden el error aleatorio del sujeto con las diferencias intra e inter observadores. 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 la conducta del sujeto en las dos ocasiones. Estos métodos por tanto 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 (Anguera y Blanco, 1984, 1988; Blanco y Anguera, 1984). 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 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 (Blanco 1986a, 1986b, 1986c).

La teoría de la generalizabilidad, que vamos a presentar, ha sido concebida justamente por sus autores (Cronbach, Rajaratnam, & Gleser, 1963) con el fin de unificar las diferentes definiciones anteriores de la fiabilidad. Como Cronbach y sus colaboradores han demostrado, estas definiciones no son contradictorias: cada una de ellas corresponde de hecho a un aspecto parcial de un modelo más general que tiene en cuenta el conjunto de todas las fuentes de variación que afectan a los resultados observados.

Gracias al concepto estadístico de muestreo de fuentes de variación múltiples, Cronbach et al. (1972) han podido tratar cada característica de la situación de observación (por ejemplo, personalidad del observador, estado subjetivo del sujeto de observación, característica de la categoría de registro, etc.) como una faceta de un diseño de observación sistemática. Aplicando las técnicas de análisis de la variancia, podremos cuantificar la importancia de cada fuente de variación.

Es posible entonces definir la puntuación verdadera como la esperanza matemática de todas las observaciones posibles, y el error como una fluctuación muestral correspondiente a la extracción aleatoria de ciertos niveles de las facetas consideradas (elección de determinados observadores, de determinados momentos, de diversos lugares,...) La teoría estadística puede decirnos el intervalo en que se encontrará la puntuación verdadera cuando utilicemos un tipo de muestreo y qué progresos conseguiremos si seleccionamos de otra forma las muestras (Suen, Lee, & Owen, in press).

La teoría de la generalizabilidad nos ofrece así un marco más satisfactorio que los conceptos anteriores para buscar las estimaciones de fiabilidad y de márgenes de error, ya que es suficientemente globalizadora como para adaptarse a las condiciones particulares de cada objeto de medida (Blanco, Losada y Anguera, 1991a, 1991b). Este carácter globalizador, sin embargo, no ha sido puesto de manifiesto por los autores de esta teoría, quienes la han formulado casi exclusivamente en términos apropiados a contextos psicométricos, en su obra de base de 1972, *The Dependability of Behavioral Measurement: Theory of Generalizability for scores and profiles* (Cronbach, Gleser, Nanda, & Rajaratnam, 1972). Las extensiones sucesivas de este enfoque ha planteado nuevas categorías de problemas, que han sido propuestos y resueltos por Cardinet & Tourneur (1985), Mitchell(1979) y Smith & Tecter (1982). Apoyándose en el hecho de que ninguno de los factores incluidos en un diseño de observación tiene una primacía particular y que, por ejemplo, la *diferenciación* de las dificultades de los observadores podría ser en ciertos casos tan importantes como las de los individuos, estos autores han podido abordar una serie de casos particulares, no considerados por Cronbach et al. (1972), que requieren un nuevo marco conceptual y procedimientos más generales de cálculo.

TEORIA DE LA GENERALIZABILIDAD: CONCEPTOS BASICOS

Las nociones clásicas de puntuación verdadera y del error van a ser reemplazadas por conceptos más en armonía con desarrollos estadísticos modernos.

Sea $X_{(ic)}$ el valor observado del individuo (i) en la categoría (c). Imaginemos que el individuo (i) sea evaluado a través de todas las posibles 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 individuos de una misma población son evaluados en una misma categoría (c), obtendremos una medida $\mu(c)$, que constituirá la *puntuación universo* ligada a la categoría (c) para todos los individuos de la población. Es decir, habríamos calculado la ocurrencia de esta categoría a partir de los individuos de la población. Utilizaremos el término *población* para los objetos de medida y el término *universo* para las condiciones de evaluación.

Si calculamos la media para todas las $\mu_{(c)}$ o para todos los $\mu_{(i)}$ obtendremos la media general μ de todas las categorías (c) del universo (que miden dicha competencia) aplicadas a todos los individuos de la población. Dicha media indicará el nivel medio de posesión de esa competencia en la población de individuos evaluados. Luego,

$$X_{ic}=\mu+\mu(i)+\mu(c)+\mu(ic)+\varepsilon \qquad (1)$$

Para toda observación correspondiente a un individuo (i) y a una categoría (c) tenemos la siguiente igualdad:

$$\begin{aligned} X_{(ic)} = & \mu \ (\textit{media general}) \\ & +\mu_{(i)} - \mu \ (\textit{efecto del individuo}) \\ & +\mu_{(c)} - \mu \ (\textit{efecto de la categoría}) \\ & +X_{(ic)} - \mu_{(i)} - \mu_{(c)} + \mu \ (\textit{efecto residual}) \end{aligned} \qquad (2)$$

Sus respectivos componentes de variancia vendrían expresados mediante la siguiente igualdad:

$$\begin{aligned} \sigma^2 X_{(ic)} = & \\ & +\sigma^2_{(i)} \\ & +\sigma^2_{(c)} \\ & +\sigma^2_{(ic,e)} \end{aligned} \qquad (3)$$

Matemáticamente la ecuación no es más que una tautología. Por tanto, la ecuación divide la puntuación observada $X_{(ic)}$ en sus componentes que representan los efectos hipotéticos.

De la misma forma, el interés ligado a las puntuaciones observadas no existe más que en la medida en que esas puntuaciones son representativas de un conjunto de puntuaciones similares. 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 que se obtendrían si se efectuasen las observaciones en todas las condiciones posibles.

El término *faceta* lo introduce Cronbach 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 puntuados los sujetos observados van a variar de un individuo a otro según el observador, lo que constituye una fuente de variación importante. La faceta "categorías" se define entonces como el conjunto de categorías posibles, a través de las cuales el observador elegirá un cierto número de ellas.

Cada una de las manifestaciones posibles de una *faceta* (cada observador, cada individuo, cada sesión, cada método de registro de datos, cada instrumento, etc.), ya que cada elemento del conjunto constituyen la *faceta*, será designada como un *nivel* de la faceta. Así el observador nº2 será considerado como un *nivel* de la faceta "observadores". En la teoría G (generalizabilidad) las facetas serán simbolizadas por letras mayúsculas.

La *puntuación universo* de un individuo (i), que es el dato ideal, representa la media de las puntuaciones del individuo (i), calculada sobre todas las observaciones posibles. O bien utilizamos la puntuación observada, o bien 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). La *generalizabilidad* es por tanto el grado por el cual podemos generalizar un resultado obtenido en unas condiciones particulares a un valor teórico buscado. El coeficiente de generalizabilidad trata de estimar en qué medida se puede generalizar a partir de la media observada en esas condiciones, a la media de todas las observaciones posibles.

El *universo de generalización* es el conjunto de condiciones a las que se quiere generalizar los resultados observados en esas condiciones particulares. Ello resulta eventualmente de la elección de un subconjunto de condiciones "admisibles" en el conjunto original de todas las condiciones posibles.

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 seleccionará un nuevo plan o diseño 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)* u *optimización* que aprovechará las informaciones obtenidas en el estudio G.

En principio, un estudio G implica calcular la parte de la variancia total que es atribuible a las diversas facetas y a sus interacciones. El conocimiento de estos valores permite optimizar la medida para los estudios D posteriores: se generalizará sobre aquellas facetas en donde se puede reducir la variabilidad en las muestras, pero se mantendrán fijas otras facetas en las que el efecto es demasiado importante o difícilmente reducible.

FASES DE LA TEORÍA DE LA GENERALIZABILIDAD

La demarcación que vamos a llevar a cabo introduce una distinción entre las fases del análisis de la variancia y las que se fundamentan en los conceptos de la teoría de la generalizabilidad. El modelo del análisis de la variancia tiene en cuenta las observaciones en las que se supone la existencia de fuentes de variancia. Permite precisar la importancia de cada una de estas fuentes de variación, atribuyéndoles una porción de la variancia total. En este modelo, nada evoca la distinción entre puntuación verdadera y del error. Todas las fuentes de variancia son necesarias en una descripción correcta y completa de la realidad observada.

Todo este desarrollo lo dividiremos en cuatro fases. Las dos primeras tienen su fundamentación en el análisis de la variancia, mientras que las fases tercera y cuarta desarrollan los conceptos que son propios a la teoría de la generalizabilidad (Brennan, 1980, 1983; Cardinet, 1987; Cardinet & Tourneur, 1985; Cardinet, Tourneur, & Allal, 1976, 1981; Shavelson & Webb, 1991).

La *primera* fase es puramente descriptiva: se identifican y organizan los datos en un *Plan de observación*. Se eligen las *facetas* a tener en cuenta y se precisan las interrelaciones entre las facetas estudiadas. Se decide el número de *niveles* muestreados en cada faceta. Se utiliza el análisis de la variancia con el fin de calcular el *cuadrado medio* de cada fuente de variación del plan utilizado. También se habrá de completar la tabla habitual de fuentes de variación del análisis de la variancia.

En las formulaciones típicas de la teoría de la generalizabilidad (Cronbach et al., 1972) todas las fuentes de variación no son tratadas de igual forma. Las variaciones que corresponden a las diferencias entre individuos son consideradas como la variancia verdadera, todas las otras fuentes de variación contribuyen a la variancia del error, disminuyendo la generalizabilidad de las puntuaciones observadas de los individuos. En la formulación de Cardinet & Tourneur (1985), el plan de observación es considerado como *simétrico*, es decir que no se hace ninguna distinción de naturaleza entre las diferentes fuentes de variación. El *objeto de medida* podrá variar, ya que los mismos datos podrán ser utilizados en los análisis sucesivos, o cada faceta podrá ser seleccionada independientemente como objeto de medida. La formulación del plan de observación determina simplemente la forma en que se deben calcular y diseñar las sumas de cuadrados y los cuadrados medios.

En la *segunda fase* del desarrollo de un análisis de generalizabilidad, la elección de un modelo de *estimación* apropiado (ya sea de efectos aleatorios o mixtos) está determinado por el modo de muestrear los niveles de cada faceta. Siguiendo la terminología del análisis de la variancia, diremos que una faceta es *aleatoria* si una muestra aleatoria simple de niveles observados se extrae de un conjunto infinito (o hipotéticamente infinito) de niveles *admisibles*. Llegados a este punto, sería útil distinguir los *niveles observados*, definidos por el plan de observación, y los *niveles admisibles*, definidos por el Plan de Estimación (desarrollado en esta segunda fase). 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 está simbolizado por N, de igual forma. En el caso de que una faceta sea *fija* $N_{(i)} = n_{(i)}$.

Consideraremos que una faceta es *fija* si los niveles admisibles son representados de manera exhaustiva en el plan de observación, es decir 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 muestreo aleatorio a partir de una población o universo finito de niveles. En tal caso, hablaremos de que la faceta es *aleatoria finita*.

En la *tercera fase*, se introducen los conceptos de la teoría de la generalizabilidad, con el fin de analizar las propiedades de uno o más Planes de Medida. Esta fase sirve para precisar la intención de medida y también para especificar

qué faceta o facetas constituyen el objeto de estudio privilegiado. Esta intención de medida crea una disimetría entre las facetas, ya que unas van jugar el papel de fuentes de variancias deseables, mientras que las otras serán fuentes de fluctuaciones aleatorias, es decir fuentes de error.

Los *objetos de medidas* admisibles constituyen la población objeto de estudio y los *instrumentos de medida* (las condiciones de observación en terminología de Cronbach) constituyen el universo de generalización. Los primeros se sitúan en el aspecto de la DIFERENCIACION, ya que la variancia verdadera proviene de las diferencias entre objetos de estudio. Los segundos se sitúan en el aspecto de la INSTRUMENTACION, puesto que las condiciones de medida son como los instrumentos o medios de esta medida. Se han seleccionado los términos *diferenciación* e *instrumentación* porque se corresponden bien con las operaciones fundamentales del análisis de la generalizabilidad, 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 la elección de los instrumentos utilizados en la medida (Cardinet, Tourneur & Allal, 1976, 1981). Así, nuestro Plan de Medida 1 (Tabla 2), definido con la siguiente estructura *ICS/O*, significa que los individuos (I), las categorías (C) y las sesiones (S) constituirán las fecetas objeto de medida (*diferenciación*), mientras que los observadores (O) serán el instrumento de medida (*instrumentación*).

Si aplicamos a esta tercera fase el principio de *simetría*, es decir tomando como objeto de estudio cada una de las facetas, podemos ver que se puede atribuir toda faceta ya sea a la diferenciación, ya sea a la instrumentación. Esta imputación se realiza independientemente del modo de distribución de los niveles de las facetas (Tabla 2). De esta forma, se tendrán en cuenta cuatro tipos de facetas: las facetas de diferenciación que son *aleatorias* (infinitas o finitas) o las que son *fijas*, y las facetas de instrumentación que son *aleatorias* (infinitas o finitas— o las que son *fijas*. En el estudio empírico que presentamos no se ha llevado a cabo un Plan de Estimación con facetas fijas y, por tanto, ningún Plan de Medida refleja la utilización de facetas fijas.

En la formulación original de la teoría de la generalizabilidad, la *diferenciación* del plan de medida está reducida a una sola dimensión compuesta de una única faceta: los individuos. Por el contrario, generalmente se incluyen muchas facetas de *instrumentación*, que definen todas las condiciones admisibles de la medida de los objetos de estudio. Cardinet, Tourneur, & Allal (1976,1981) proponen utilizar un concepto paralelo en la diferenciación: puesto que la población estudiada, ya sea individuos u otros objetos de estudio, está estratificada en función de un cierto número de criterios, el aspecto de *diferenciación* en un plan cruzado estará formado por el producto cartesiano de todas las facetas de diferenciación. Así, en un plan cruzado completo (como el que presentamos empíricamente en este trabajo), el conjunto de observaciones admisibles de un estudio de generalizabilidad está definido por el producto cartesiano del conjunto de

objetos admisibles de estudio, cruzando con el conjunto de condiciones admisibles de medida.

En la *cuarta fase*, las informaciones obtenidas en los análisis precedentes se utilizan para identificar la mejor adecuación posible en los procedimientos de medida. Ello nos conducirá posiblemente a la elección de otra disposición mejor adaptada a ciertas condiciones de decisión. En terminología de Cronbach et al. (1972) estaríamos hablando de un estudio de *decisión* (D), aunque en la nueva terminología de Cardinet & Tourneur (1985) se trataría del *plan de optimización*, en base a que esta fase se aplica tanto a las situaciones de medida orientadas hacia una decisión como a las áreas de investigación orientadas hacia una conclusión. El plan de optimización se establece en base a las informaciones procedentes de la tercera fase, modificando el Plan de Observación o el de Estimación o el de Medida o varios de ellos a la vez.

Hay que resaltar finalmente que estas cuatro fases indican el orden en el que deben ser aplicados los procedimientos de análisis y de estimación, una vez que los datos han sido agrupados para un estudio de generalizabilidad (Duquesne, 1986), pero no describen el orden en el que las consideraciones conceptuales son tratadas cuando se planifica un estudio de generalizabilidad.

FUNDAMENTOS TEÓRICOS DE LA ESTIMACION DE LA PRECISIÓN

La obra de Cardinet & Tourneur (1985) nos suministra procedimientos generales de cálculo, aunque otros métodos alternativos pueden encontrarse en la propia obra de Cronbach et al. (1972), en Brennan (1983), en Marcoulides (1989) y en Shavelson & Webb (1991). Basándonos en el trabajo de Cardinet, Tourneur & Allal (1976), se pueden situar fácilmente las fuentes de error a controlar a partir de un gráfico que represente todas las *facetas* del diseño.

Supongamos por ejemplo que diferentes observadores (O) han evaluado a diversos individuos (I) en un sistema de categorías (C) a través de una serie de sesiones de observación (S). De esta forma, se puede construir la Figura 1 (correspondiente al Plan de Medida 1 *ICS/O*, en la que las facetas de los *objetos de estudio* (individuos, categorías y sesiones) están representadas mediante líneas horizontales y las facetas de los *instrumentos de medida* (observadores) mediante líneas verticales.

Las fuentes de error que afectan a la comparación de resultados de los individuos (I) evaluados en un sistema de categorías (C) a través de diversas sesiones (S) se sitúan en la intersección (líneas cruzadas) de objetos de estudio e instrumento de medida. Se trata de las variancias de interacción: IO, ISO, IOCS, IOC, CO, CSO, SO, que el análisis de variancia permite estimar para todo el conjunto de datos correspondientes a este diseño de investigación. Dado que los observadores (O) son el *instrumento de medida* se trata de todas las interacciones de primer y segundo orden de todas las fuentes de variación que incluyen (O), es decir que afectarán a la fiabilidad intraobservadores. A partir de estas estimaciones se

puede comprobar en particular si es o no más útil aumentar el número de observadores en estudios posteriores.

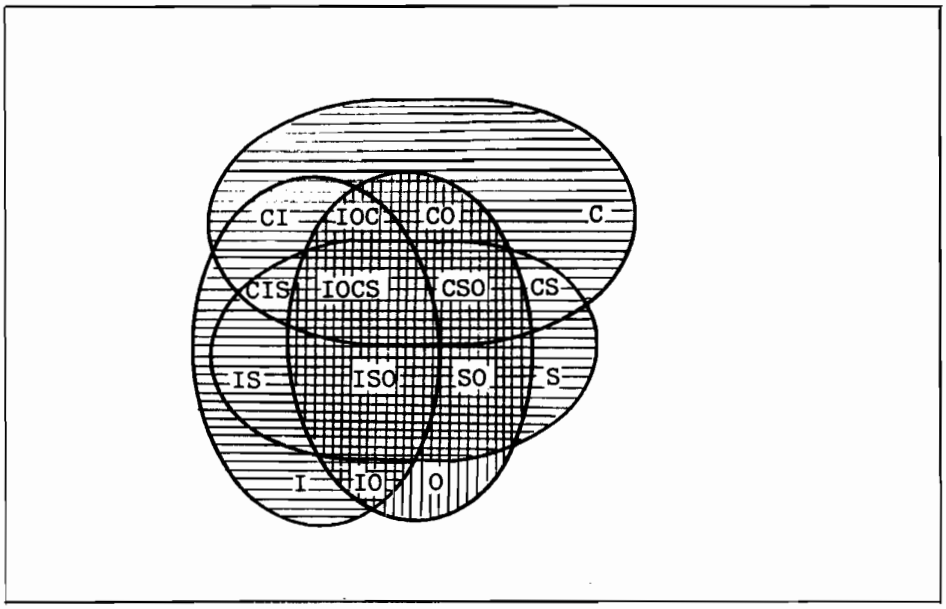

Figura 1. Representación gráfica del Plan de Medida 1 (ICS/O).

La Figura 2 (correspondiente al Plan de Medida 6 *ISO/C*), en donde el *instrumento de medida* son las categorías (C), permite a partir de dichas estimaciones comprobar la homogeneidad del sistema de categorías y si es o no más útil aumentar el número de niveles de las mismas. En tal caso, las fuentes de error se sitúan igualmente en la intersección (líneas cruzadas) de *objetos de estudio* e *instrumento de medida*: CI, CIS, IOC, IOCS, CO, CSO, CS, es decir las variancias de interacción estimadas para el conjunto de datos correspondientes a este Plan de Medida que incluyen (C) y que afectarán al *instrumento de medida* categorías.

La Figura 3 nos presenta el Plan de Medida 8 IO/SC. Las fuentes de error que afectan a la comparación de resultados de los individuos (I) y observadores (O) conjuntamente se sitúan en la intersección (líneas cruzadas) de *objetos de estudio* e *instrumentos de medida*. Se trata de las variancias de interacción: IS, CIS, CI, ISO, IOCS, IOC, SO, CSO, CO (es decir, que incluyen individuos [I] y/o observadores [O]), que el análisis de la variancia permite estimar para todo el conjunto de datos correspondientes a este diseño. Se puede comprobar si un sistema de categorías utilizado en cinco sesiones ha sido capaz de clasificar a los individuos evaluados por dos observadores. Esta verificación nos permite detectar si es necesario aumentar o disminuir el número de sesiones y/o categorías conjuntamente.

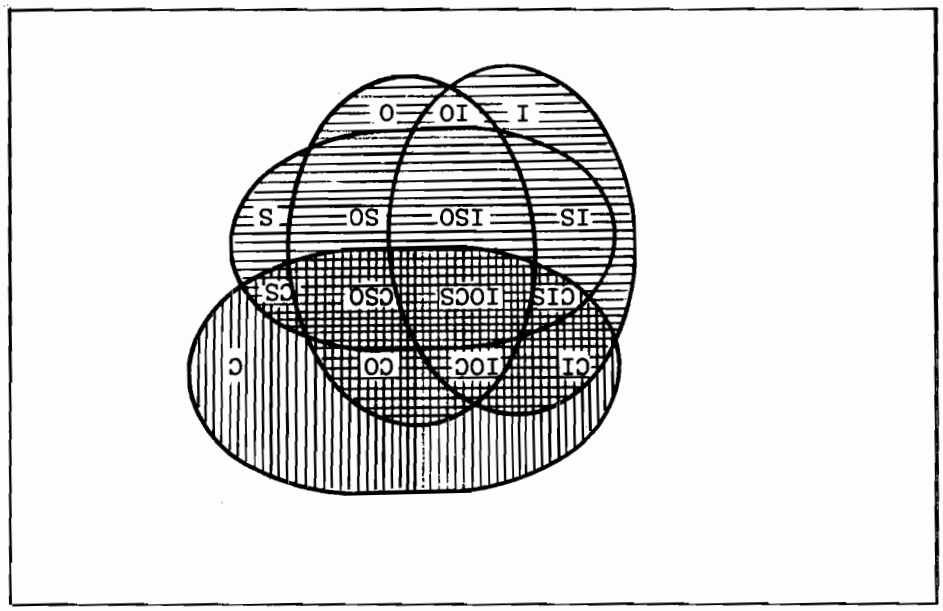

Figura 2. Representación gráfica del Plan de Medida 6 (ISO/C).

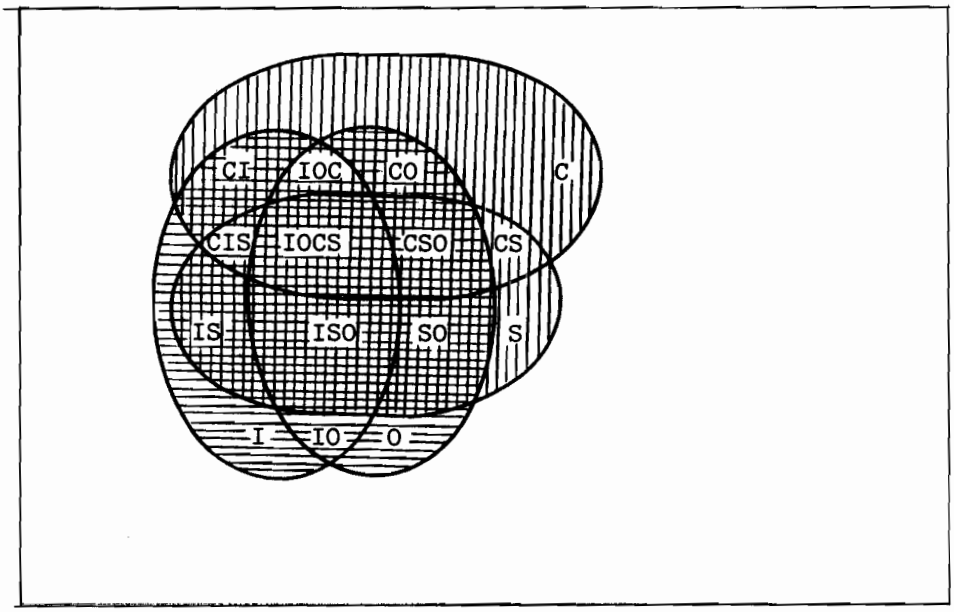

Figura 3. Representación gráfica del Plan de Medida 8 (IO/SC).

Finalmente, las fuentes de error que afectan a la *diferenciación de sesiones y observadores (Plan de Medida 9 SO/IC*, representado mediante diagramas en la Figura 4) serán aquellas interacciones que contienen individuos y/o categorías: IS, CIS, IO, ISO, IOCS, IOC, CSO, CO, CS. A partir de estas *estimaciones* se puede comprobar en particular si es o no más útil en estudios posteriores aumentar o disminuir el número de niveles en una faceta de *generalización* (individuos, categorías) a medida que aumenta o disminuye el número de niveles en la otra faceta de generalización. De esta forma, conseguiremos el número óptimo de registros que permitirán detectar diferencias significativas en los diferentes observadores evaluando en días diferentes.

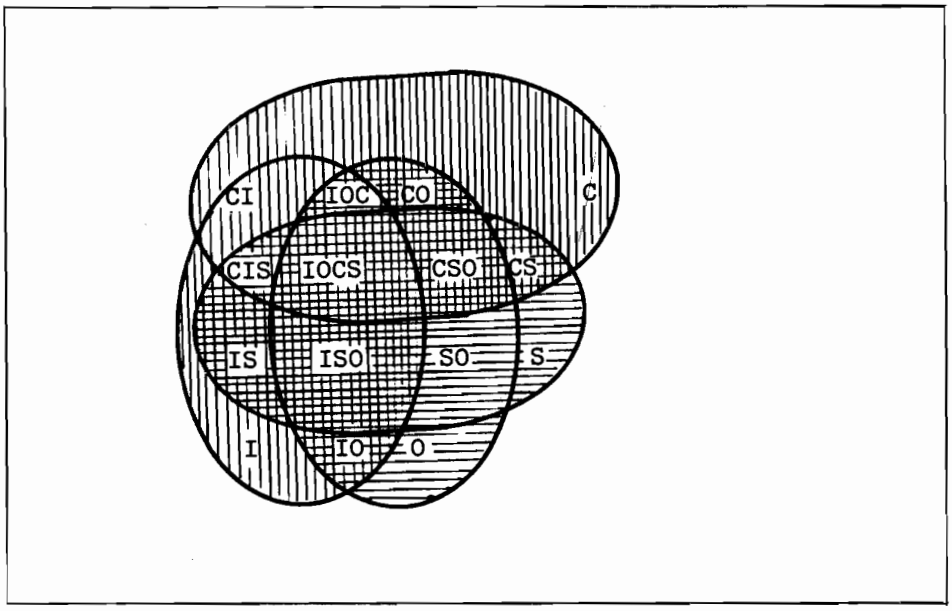

Figura 4. Representación gráfica del Plan de Medida 9 (SO/IC).

Dado el carácter abstracto, la teoría estadística puede aplicarse a cualquiera que fuese el *objeto de estudio* y es por ello que la *teoría de la generalizabilidad* constituye una teoría general de la medida (Brennan, 1983, p.12): "el modelo de medida mejor definido actualmente existente".

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 univer-*

so relativa a la variancia del error es el *coeficiente de generalizabilidad* ($E\rho^2$), que 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 (σ^2_τ) y el valor esperado de la variancia de puntuaciones observadas ($E\sigma^2_x$)

$$E\rho^2 = \frac{\sigma^2_\tau}{E\,\sigma^2_x} \tag{4}$$

o bien, dado que $E\sigma^2_x = \sigma^2_\tau + \sigma^2\delta$, donde $\sigma^2\delta$ es la variancia de error relativo

$$E\rho^2 = \frac{\sigma^2_\tau}{\sigma^2_\tau + \sigma^2\delta} \tag{5}$$

o bien, dado que $E\sigma^2_x = \sigma^2_\tau + \sigma^2\Delta$, donde $\sigma^2\Delta$ es la variancia de error absoluto

$$E\rho^2 = \frac{\sigma^2_\tau}{\sigma^2_\tau + \sigma^2\Delta} \tag{6}$$

Podemos obtener una estimación del coeficiente de generalizabilidad [$E\rho^2$] a través de estimaciones muestrales de los parámetros de la ecuación anterior

$$E\rho^2 = \frac{\sigma^2_\tau}{\sigma^2_\tau + \sigma^2\delta} \tag{7}$$

o bien

$$E\rho^2 = \frac{\sigma^2_\tau}{\sigma^2_\tau + \sigma^2\Delta} \tag{8}$$

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

La diferencia entre $\sigma^2\delta$ y $\sigma^2\Delta$ es que la primera no incluye las fuentes de variancia común a cada individuo (objeto de medida) mientras que la segunda 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 observadores (I x O), por ejemplo, $\sigma^2_i / [\sigma^2_i + \sigma^2\delta]$ es algebraicamente (pero no conceptualmente) idéntico al α de Cronbach (Brennan, 1983).

Desarrollo del diseño: Planes de Medida

Para ilustrar nuestro análisis de generalizabilidad utilizaremos un diseño totalmente cruzado correpondiente al Plan de Observación I x O x C x S. Un total de ocho individuos (I) —cuatro de sexo femenino y cuatro de masculino, aunque sin llevar a cabo en el diseño la anidación de esta faceta— han sido evaluados por dos

observadores (O) en un sistema de siete categorías (C) —donde A representaría estados agresivos de la conducta infantil y G estados de conducta prosocial, con una gradación de la intensidad de estos dos estados extremos—, a través de cinco sesiones de observación diferentes (S), realizadas en la misma situación interactiva de juego en el patio de un colegio, pero en días diferentes. Dicho plan de observación comprenderá 4 facetas: los individuos (I, que son 8), los observadores (O, igual a 2), las diferentes categorías del sistema (C, un total de 7) y las sesiones realizadas en cinco días diferentes (S, un total de 5). Los valores figuran en una tabla cruzada de doble entrada I x O x C x S que contiene 8 x 2 x 7 x 5 = 560 casos (Tabla 1). Los datos observacionales corresponden a un registro continuo, que indica la frecuencia de ocurrencia de cada una de las categorías del sistema en el tiempo total de duración de cada una de las sesiones de observación.

Tabla 1.– Matriz de datos del diseño: Ocho individuos (I) han sido evaluados por dos observadores (O) en un sistema de siete categorías (C) a través de cinco sesiones realizadas en días diferentes (S).

		A	B	C	D	B	F	G
		12345	12345	12345	12345	12345	12345	12345
1	1	01100	00110	11111	13432	44322	56224	89778
	2	10100	00111	11111	13433	34322	46225	89777
2	1	10111	11011	11111	01211	25213	44564	69558
	2	10111	11000	11111	01111	25212	44464	69557
3	1	11212	00110	12110	22321	24121	26425	78666
	2	11212	00011	12210	22221	24021	26435	78766
4	1	00121	10110	01121	14132	05202	33534	25646
	2	00011	10101	01121	14132	15212	33434	35646
5	1	11111	22212	22203	12332	22531	16312	37535
	2	11111	12212	12202	12332	22431	16412	35735
6	1	21222	21221	22212	02412	15642	25223	33343
	2	21122	21221	22212	02312	14542	25224	33343
7	1	01000	01120	01111	21331	55465	33342	23535
	2	01001	01120	01111	21331	55366	32342	23645
8	1	11110	01111	23121	01212	16241	45654	87876
	2	11110	01112	22112	01222	15242	45654	87876

Dado que cada una de las facetas puede estimarse de forma *infinita, finita o fija* (aunque en nuestro caso no hemos previsto en ninguno de los diseños la inclusión de una faceta fija) y puesto que cada una de ellas puede entrar a formar parte de la *diferenciación* o de la *instrumentación*, según el principio de "simetría de las facetas" (Cardinet, Tourneur, & Allal, 1976; Cardinet, 1987), son numerosos los planes de medida que pueden resultar. Desde nuestra perspectiva hemos seleccionado 14 Planes de Medida diferentes, intercambiando la posición de las facetas, ya sea individual, ya sea conjuntamente, cuando están situadas en el aspecto de la *diferenciación* o en el de la *instrumentación*. Cada uno de estos posibles Planes de Medida nos aportará probablemente unos resultados y un análisis diferente en función de la información que suministran (ver Tabla 2). Todos los Planes de Medida han sido analizados (ver Tabla 5) en el caso de contar con las 3 primeras sesiones de observación y en el caso de añadir dos nuevas sesiones según la información suministrada en la Tabla 1. De la misma forma, se ha comparado la información en 3 y 5 sesiones con el mismo Plan de Estimación y en 5 sesiones con difernte Plan de Estimación.

Para llevar a cabo posteriormente las optimizaciones (es decir, un estudio de decisión *D*, en terminología de Cronbach) se han seleccionado 4 Planes de Medida en que la faceta categorías, conjuntamente con otras facetas, ha sido considerada en los cuatro planes objeto de estudio o diferenciación, para comprobar la precisión en la generalización a observadores, sesiones e individuos, ya sea individual ya sea conjuntamente (ver Tabla 6). Luego hemos seleccionado otros 4 Planes de Medida en que la faceta categorías, individual o conjuntamente con otras facetas, ha sido considerada como instrumentación o instrumento de medida, con el fin de comprobar la consistencia o validez de las diferentes categorías del sistema, la generalización a individuos y categorías conjuntamente, las diferencias que existen entre las sesiones y si los observadores registran o no de una forma similar (ver Tabla 7). La optimización de los diferentes Planes de Medida siempre se ha realizado en función de la información obtenida en el Plan de Estimación de 5 sesiones, donde $N_i = N_c = N_s = \infty$, $N_o = 20$ (ver Tabla 4).

El Plan de Medida 1 (Tabla 2) intenta diferenciar a individuos (I), categorías (C) y sesiones (S). Dichas facetas constituirán la *diferenciación* de este plan. El estudio *G* evaluará a los observadores con el fin de estimar la generalización a los mismos. De esta forma, los observadores (O) conformarán la faceta de generalización. Dicho plan nos proporcionará información sobre una de las posibles fuentes de variación, los observadores, y por tanto nos ofrecerá la evaluación de la fiabilidad interobservadores. Es decir, obtendríamos los resultados del *coeficiente de fiabilidad interobservadores* y, a la vez, podríamos determinar si los resultados del estudio pueden ser generalizados con precisión a un número mayor de observadores extraídos aleatoriamente de la población origen de donde provienen.

Tabla 2.- INFORMACIÓN que podemos obtener de cada uno de los Planes de Medida.

1	ICS/O	Fiabilidad interobservadores
2	IC/SO	Los dos observadores, a través de diferentes sesiones, ¿han sido capaces de diferenciar a los individuos en las categorías?
3	I/CSO	¿Es posible diferenciar unos individuos de otros?
4	C/ISO	¿Son homogéneas las diferentes categorías?
5	CS/IO	¿Podemos generalizar nuestros resultados con precisión a los individuos y observadores?.
6	ISO/C	Validez del sistema de categorías.
7	IS/OC	Los dos observadores, a través de las siete categorías, ¿han sido capaces de diferenciar a los individuos por sesiones?.
8	IO/SC	¿Son diferentes los individuos, evaluados por dos observadores, cuando se utilizan diferentes categorías y distintas sesiones?
9	SO/IC	¿Es posible generalizar con precisión, en futuras investigaciones, a los individuos evaluados a través de un sistema de siete categorías?
10	ICO/S	Relación costo-beneficio acerca del número de sesiones utilizadas o a utilizar en futuras investigaciones.
11	COS/I	¿Han sido suficientes los individuos?. ¿Podemos generalizar con precisión a otros individuos extraídos de la misma población?
12	S/IOC	¿Se diferencian unas sesiones de otras?
13	O/ICS	¿Son diferentes los observadores?
14	OC/IS	¿Cuál es el número de sesiones e individuos, conjuntamente, necesarios para generalizar con precisión?

El plan 2 *diferencia* a individuos y categorías, mientras que sesiones y observadores constituirán las facetas de *instrumentación*. Es decir, los observadores a través de diferentes sesiones, ¿serán capaces de clasificar a los diferentes individuos en las diferentes categorías? Si el valor de la variancia de diferenciación fuera pequeño, probablemente sí; y en consecuencia generalizaríamos con mayor precisión a sesiones y observadores. En nuestro caso, dado que la variancia de diferenciación no es muy grande, conseguimos una buena precisión en la generalización a sesiones y observadores.

En el plan 3 se intenta *diferenciar* sólo a los individuos, ya que el investigador tiene unos objetivos diferentes y está interesado en la variación de los individuos, independientemente de las categorías o sesiones, como había ocurrido en los dos planes anteriores. Así, los individuos (I) constituirán el *objeto de estudio*, mientras que categorías, sesiones y observadores serán las facetas de *generalización*. La faceta observadores (O), al ser aleatoria finita, $N_O = 20$ ó $N_O = 10$, probablemente reducirá el error de generalización a categorías y sesiones.

En el plan 4 se *diferencian* las categorías (C), es decir respondería a la pregunta de si son o no homogéneas. Generalizaríamos a las otras tres facetas: individuos, sesiones y observadores.

El plan 5 *diferencia* a categorías (C) y sesiones (S), mientras que individuos (I) y observadores (O) constituirán las facetas de *generalización*. El plan nos puede proporcionar información acerca del número ideal de individuos y observadores para generalizar aún con mayor precisión.

En el plan 6, individuo, sesiones y observadores constituyen las facetas de *diferenciación*, mientras que las categorías (C) conformarán la faceta de *instrumentación* o instrumento de medida. Ello nos permitirá determinar la consistencia de las diferentes categorías utilizadas y comprobar si las mismas sirven para clasificar a los individuos, evaluados por dos observadores, en las diferentes sesiones. En este caso el investigador está interesado en generalizar al universo infinito de categorías. La inclusión conjunta de individuos, sesiones y observadores, probablemente incremente el error de diferenciación. Este hecho viene a demostrar que, en general, podríamos haber anidado una faceta con otra (por ejemplo anidar a los individuos en las sesiones o clasificar a los individuos según el sexo), ya que disminuiría el error de diferenciación y probablemente el costo en el número de niveles de las facetas. De esta forma, las facetas así clasificadas o estratificadas nos permitirían determinar si dichas anidaciones afectan la variancia de los objetos que van a ser diferenciados (Rentz, 1987). Si los componentes de variancia fueran relativamente pequeños, deberán ser eliminados en estudios posteriores de optimización (Cardinet, Tourneur, & Allal, 1981).

El plan 7 responde a la pregunta de si observadores y categorías clasifican a los individuos por sesiones. Dado que la faceta observadores es aleatoria finita, es posible que reduzca algo el error de la generalización.

En el plan 8 se consideran conjuntamente como instrumentos de medida de la investigación a sesiones y categorías. El investigador está interesado en com-

probar si las categorías a través de las diferentes sesiones han sido capaces de diferenciar con precisión nuestro objeto de estudio (individuos y observadores); en consecuencia, si podremos generalizar con precisión a sesiones y categorías conjuntamente.

En el plan 9 se intenta verificar si, en futuras investigaciones, podemos generalizar con precisión a los individuos evaluados a través de un sistema de siete categorías, y por tanto de si hemos de aumentar el número de individuos o categorías. Al mismo tiempo, nos permite determinar si los observadores han sido clasificados en las diferentes sesiones.

El plan 10 es un claro ejemplo de valoración de la relación costo-beneficio, ya que un incremento del número de sesiones incide relativamente en el presupuesto de la investigación. Por tanto nos permite determinar si generalizaríamos con precisión a ese número de sesiones o hemos de aumentar las mismas para reducir los posibles errores de medida.

El plan 11 lleva a cabo una evaluación general de la estructura del diseño, ya que nos dice cómo las diferentes fuentes de variación pueden afectar a los individuos, que en definitiva constituyen el objeto esencial de la investigación en ciencias del comportamiento. Es decir, si con un número mayor de individuos reducimos el error de generalización a los mismos. Dado que los individuos son la faceta de *generalización* podemos fijar el número ideal de individuos que serán necesarios extraer aleatoriamente de la población origen de donde provienen. Probablemente, si la generalización fuera excelente, sería conveniente anidar esta faceta clasificando a los individuos según el sexo. De esta forma reduciríamos al mínimo el error de generalización y nuestra precisión de estimación sería mucho mayor. Al mismo tiempo lograríamos reducir el número de niveles de la faceta.

El plan 12 nos presenta como único objeto de estudio a las sesiones (S), es decir si se diferencian unas de otras. El investigador está interesado en calcular la precisión de la estimación cuando quiere generalizar a un conjunto aleatorio de individuos, observadores y categorías. Si el valor de precisión en la generalización fuera pequeño, querrá decir que no existen diferencias significativas entre las tres o las cinco sesiones y que por tanto la información registrada en las mismas es similar o igual.

El plan 13 tiene como objeto de estudio o *diferenciación* a los observadores (O). Si el valor de precisión en la generalización fuera alto nos indicará que las diferencias entre ellos es muy significativa y que registran información diferenciada para cada individuo, categoría y sesión. En contrapartida, si el valor fuera pequeño, querrá decir que las diferencias interobservadores no son significativas y que por tanto apenas han cometido errores y registran una información más o menos similar (sería el Plan de Medida opuesto al número 1).

Finalmente, el Plan de Medida 14 nos permitirá determinar también la relación costo-beneficio de individuos (I) y sesiones (S) conjuntamente. El error de medida de una faceta u otra nos permitirá aumentar o disminuir el número de niveles en una u otra dependiendo de la fuente de variación que mayor error apor-

te. En consecuencia, ¿cuántos individuos y sesiones, dependiendo una de otra, serán necesarios para generalizar con precisión en investigaciones futuras, donde los individuos han sido seleccionados aleatoriamente de la población origen de donde provienen?

RESULTADOS

A continuación se presentan los *componentes de variancia* y su respectiva contribución en porcentajes correspondientes al Plan de Observación I x O x C x S, pero llevando a cabo cinco posibles estimaciones, según los objetivos de nuestro estudio (Tabla 4). En el *primer bloque* se han realizado tres estimaciones diferentes para detectar la variabilidad de las cinco sesiones: una de ellas totalmente aleatoria; otra mixta, sin variar la aleatoriedad en la que deberán ser seleccionados los individuos, las categorías y las sesiones, pero sí la variabilidad que podría producir la estimación aleatoria de los observadores (O) que es considerada una faceta aleatoria finita $N_O = 20$; y la tercera estimación, también mixta, en la que no se modifica la aleatoriedad de los individuos, pero se reduce la estimación en el número de observadores $N_O = 10$ y se consideran también finitas las facetas categorías y sesiones, N_C 0 100 y $N_S = 10$. En el *segundo bloque*, tan sólo se han considerado las 3 primeras sesiones de observaciones mostradas en la matriz de datos del diseño de la Tabla 1; se presentan en tal caso dos estimaciones, una de ellas totalmente aleatoria y la otra mixta (teniendo en cuenta de nuevo la variabilidad que podría producirse en el caso de que sólo la faceta observadores (O) sea aleatoria finita $N_O = 20$. Dicha información ha sido obtenida a través del programa BMDP 8V del paquete de programas BMDP en su última revisión PC90 (Dixon, Brown, Engelman, & Jennrich, 1990). Todo ello en base a la información inicial suministrada por los cuadrados medios de la tabla resumen del análisis de la variancia (Tabla 3), donde podemos advertir ya gran variabilidad en la faceta categorías (C) y sesiones (S), individualmente, y en sus interacciones de primer y segundo orden con las otras facetas.

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. En el *primer bloque* de la Tabla 4, $N_i = N_O = N_C = N_S = \infty$, σe nos ofrecen porcentajes de variación imperceptibles para las facetas individuos y observadores, así como para las demás interacciones de primer orden en las que están implicadas estas facetas. No observamos tampoco ninguna otra variación, con respecto a estas facetas, en las otras dos estimaciones mixtas. Estos datos no serían muy deseable si alguna de estas facetas, individual o conjuntamente, constituyeran las facetas de *diferenciación*, ya que impedirían generalizar los resultados a las demás facetas consideradas como *instrumentación*. En definitiva, las diferencias no son significativas para ambas facetas y no permiten por tanto una clasifi-

cación de los individuos y/o de los observadores. Ahora bien, si estas facetas constituyeran la *generalización*, ya sea individual ya sea conjuntamente, obtendríamos excelentes resultados de generalización a las mismas. Incluso nos permitirían generalizar con un número menor de niveles en las facetas.

Tabla 3.- Cuadro resumen del análisis de la variancia del Plan de Observación I x O x C x S.

FUENTES DE VARIACION	SUMA DE CUADRADOS	G.L.	CUADRADOS MEDIOS
INDIVIDUOS (I)	27.241	7	3.892
OBSERVADORES (O)	0.088	1	0.088
IO	0.612	7	0.087
CATEGORIAS (C)	1432.768	6	238.795
IC	346.946	42	8.261
OC	0.275	6	0.046
IOC	3.725	42	0.089
SESIONES (S)	61.118	4	15.279
IS	44.339	28	1.584
OS	1.010	4	0.253
IOS	3.647	28	0.130
CS	74.107	24	3.088
ICS	290.036	168	1.726
OCS	4.314	24	0.180
IOCS	15.828	168	0.094

El efecto contrario puede advertirse en la estimación de los componentes de la faceta *categorías (C)*, que aportan un 62% aproximadamente de la variabilidad total del diseño de investigación en las tres estimaciones. Todas las interacciones de *C* de primer y segundo orden también aportan un porcentaje importante de variabilidad. En dicha situación, la faceta categorías aportará mucho error de medida y, por tanto, cuando se constituya en faceta de *instrumentación*, el coeficiente de generalizabilidad tenderá a ser nulo. En contrapartida, la *diferenciación* de las categorías tendrá una buena precisión, ya que las diferencias serán muy significativas y las categorías tenderán claramente a diferenciarse una de otra.

También observamos cierta variabilidad en el componente de variancia *sesiones (S)*, aunque se mantiene la misma en las tres estimaciones, incluso considerando la faceta aleatoria finita N_s 0 10. Este hecho también podemos

detectarlo en las interacciones de (S), particularmente en la interacción de segundo orden *ICS* (que aporta aproximadamente un 17% del total de la variabilidad del diseño). Por tanto, si dicha faceta es considerada como *instrumento de medida*, también aportará cierto error al diseño de medida. En tal caso, podemos predecir desde este momento que la generalización a las 5 sesiones será buena, pero no excelente. Ello significará aumentar el número de niveles de la faceta sesiones para reducir el error de medida y así generalizar aún con mayor precisión. Por tanto, de la información suministrada por los *componentes de variancia*, individualmente, podemos adelantar que el número de sesiones y categorías seleccionadas no ha sido suficiente para obtener precisión en ambas facetas cuando generalicemos a las mismas y sí han sido suficientes el número de niveles seleccionados de individuos y observadores. Posteriormente, en los planes de optimización, determinaremos cuál debe ser el aumento en categorías y sesiones para estimar con precisión en investigaciones futuras.

Tabla 4.– Componentes de variancia aleatorios y/o mixtos y su contribución en porcentajes, correspondientes al Plan de Observación I x O x C x S, en cuatro Planes de Estimación diferentes.

	$N_i = \infty$ $N_o = \infty$ $N_c = \infty$ $N_s = \infty$	%	$N_i = \infty$ $N_o = 20$ $N_c = \infty$ $N_s = \infty$	%	$N_i = \infty$ $N_o = 10$ $N_c = 100$ $N_s = 10$	%	$N_i = \infty$ $N_o = \infty$ $N_c = \infty$ $N_s = \infty$	$N_i = \infty$ $N_o = 20$ $N_c = \infty$ $N_s = \infty$	%
	5 SESIONES						3 SESIONES		
I	-0.0598	0.00	-0.0599	0.00	-0.0538	0.00	0.0017	0.0015	0.03
O	0.0000	0.00	0.0000	0.00	0.0001	0.00	0.0036	0.0034	0.07
IO	-0.0011	0.00	-0.0010	0.00	-0.0004	0.00	-0.0036	-0.0034	0.00
C	2.8663	61.82	2.8661	61.82	2.8453	62.14	3.1127	3.1127	60.52
IC	0.6540	14.10	0.6539	14.10	0.7291	15.92	0.5247	0.5241	10.19
OC	-0.0032	0.00	-0.0030	0.00	-0.019	0.00	-0.0002	-0.0002	0.00
IOC	-0.0011	0.00	-0.0011	0.00	0.074	0.16	-0.0117	-0.0111	0.00
S	0.1098	2.37	0.1098	2.37	0.0996	2.18	0.2253	0.2251	4.38
IS	-0.0128	0.00	-0.0125	0.00	-0.0036	0.00	-0.0502	-0.0499	0.00
OS	0.0007	0.01	0.0006	0.00	0.0006	0.01	-0.0025	-0.0024	0.00
IOS	0.0051	0.11	0.0049	0.11	0.0049	0.11	0.0055	0.0052	0.10
CS	0.0797	1.72	0.0803	1.73	0.0720	1.576	0.1280	0.1286	2.50
ICS	0.8161	17.60	0.8208	17.70	0.7355	16.06	1.0272	1.0325	20.07
OCS	0.0107	0.23	0.0102	0.22	0.0086	0.19	0.0110	0.0104	0.20
IOCS	0.0942	2.03	0.0895	1.93	0.0756	1.65	0.1051	0.0998	1.94

En cuanto a los demás componentes de variancia, el resto de interacciones de primer y segundo orden, nos ofrecen valores también relativamente pequeños (excepto la ya comentada interacción *ICS*). En conjunto, podemos sugerir que, si los individuos y observadores constituyeran en el plan de medida una faceta de generalización, en estudios posteriores podríamos reducir incluso el número de los mismos, sin que por ello reduzcamos la precisión en la generalización. Por el contrario, dado que las categorías (C) contribuyen al error de una manera significativa, en el caso de que las categorías fueran la faceta de generalización en el plan de medida, necesitaríamos aumentar el número de las mismas si deseamos un nivel alto de generalizabilidad.

En lo que respecta al *segundo bloque* de estimaciones (Tabla 4), que corresponde a una totalmente alcatoria y a otra mixta si hubiéramos tenido en cuenta sólo los registros de las tres primeras sesiones, podemos observar resultados similares, aunque sólo pueden interpretarse si los comparamos con el primer bloque. El Componente (I) es positivo y aporta variabilidad, algo que no ocurría con las estimaciones en cinco sesiones. El componente (C) disminuye con respecto a las cinco sesiones, es decir a mayor número de sesiones se obtiene mayor variabilidad en las categorías. Algo similar ocurre con la interacción de ambas (IC), que también disminuye notablemente. El componente (O) también aporta algo de variabilidad, mientras que en cinco sesiones era prácticamente nula; ello viene a demostrar la consistencia y fiabilidad de los observadores en un número mayor de sesiones. Y, evidentemente, la variabilidad del componente sesiones (S) *dobla* su valor, por lo que sí ha sido necesario aumentar los niveles de la faceta sesiones. Y, en general, se observa mayor variabilidad en el resto de interacciones de primer y segundo orden. En definitiva, este análisis previo de los componentes de variancia nos ha facilitado una información anticipada, que probablemente se corroborará en la fase final de optimización del diseño, y que nos facilitará posteriormente la selección de nuevos niveles en las facetas que mayor error de medida aporten y rediseñar así nuestra investigación.

La estimación de algunos *componentes de variancia* (I, IO, OC, IOC, IS, en el primer bloque; IO, OC, IOC, IS, OS, en el segundo bloque) nos ofrecen un valor negativo (Tabla 4) y, aún cuando es imposible encontrar una suma de cuadrados negativa, es debido a las fluctuaciones muestrales en la estimación de los cuadrados medios. En estos casos, tanto Cronbach et al. (1972) como Cardinet et al. (1981), proponen reemplazar el valor negativo encontrado por un valor nulo. La modificación y optimización de este diseño, para lograr mayor precisión en la generalización de los diferentes planes de medida es lo que será considerado a continuación.

Los coeficientes de generalizabilidad absolutos [Eo $^2\Delta$] y relativos [Eo$^2\delta$] de los 14 planes de medida se presentan en la Tabla 5. Estos coeficientes estiman la generalizabilidad a través de las facetas de generalización (observadores en el plan 1; sesiones y observadores en el plan 2; categorías, sesiones y observadores en el plan 3; individuos, sesiones y observadores en el plan 4; individuos y obser-

vadores en el plan 5,...). Por ejemplo, en el plan 1, que el coeficiente de generalizabilidad relativo tiene un valor más próximo a la *unidad*, podría interpretarse como la correlación entre un conjunto de 100 puntuaciones y otro conjunto de otras 100 puntuaciones obtenidas en otro conjunto a través de los registros de dos observadores. En el plan 2, que también el coeficiente de generalizabilidad relativo aporta un valor próximo a la *unidad*, podría interpretarse como la correlación entre un conjunto de 100 registros observacionales y otro conjunto de otros 100 registros observacionales obtenidos en otro conjunto a través de los registros de dos observadores en tres o cinco sesiones diferentes. En el plan 6, que el valor del coeficiente tiene a cero, entendemos que no existe relación entre un conjunto de 100 registros y otro conjunto de otros 100 registros obtenidos en otro conjunto a través de un sistema de cinco categorías, es decir no tendremos *precisión en la generalización* (generalizabilidad en terminología de Cronbach). La magnitud del coeficiente (que varía de 0 a 1) se interpreta de la misma forma que los coeficientes de fiabilidad tradicionales.

Tabla 5.– Posibles Planes de Estimación y Medida (en tres y cinco sesiones) para ilustrar el análisis de la generalizabilidad del Plan de Observación I x O x C x S, $n_i = 8$, $n_o = 2$, $n_c = 7$, $n_s = 5$.

	3 SESIONES		5 SESIONES		5 SESIONES	
PLAN	$N_i = \infty$		$N_i = \infty$		$N_i = \infty$	
DE	$N_o = 20$		$N_o = 20$		$N_o = 10$	
MEDIDA	$N_c = \infty$		$N_c = \infty$		$N_c = 100$	
	$N_s = \infty$		$N_s = \infty$		$N_s = 10$	
	$E\rho^2_\delta$	$E\rho^2_\Delta$	$E\rho^2_\delta$	$E\rho^2_\Delta$	$E\rho^2_\delta$	$E\rho^2_\Delta$
1.- ICS/O	0.98	0.90	0.98	0.98	0.99	0.99
2.- IC/SO	0.90	0.88	0.94	0.94	0.97	0.97
3.- I/CSO	0.012	0.002	0.00	0.00	0.00	0.00
4.- C/ISO	0.95	0.95	0.96	0.95	0.96	0.95
5.- CS/IO	0.94	0.94	0.94	0.94	0.94	0.94
6.- ISO/C	0.47	0.25	0.32	0.15	0.32	0.14
7.- IS/OC	0.47	0.24	0.32	0.14	0.31	0.14
8.- IO/SC	0.036	0.007	0.0002	0.00	0.0005	0.0001
9.- SO/IC	0.84	0.51	0.788	0.197	0.79	0.19
10.- ICO/S	0.89	0.87	0.94	0.94	0.97	0.97
11.- COS/I	0.94	0.94	0.94	0.94	0.94	0.94
12.- S/IOC	0.853	0.313	0.796	0.196	0.80	0.19
13.- O/ICS	0.72	0.06	0.24	0.00	0.10	0.00
14.- OC/IS	0.95	0.93	0.95	0.95	0.96	0.95

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 buenas en los planes en que las categorías (C), individual o conjuntamente, constituyen el *instrumento de medida* (Planes 3, 6, 7, 8, 9, 12 y 13). Por el contrario, la generalización aumenta cuando el *objeto de estudio* son las categorías (C), individualmente (Plan 4) o conjuntamente con otras facetas (Planes 1, 2, 5, 10, 11 y 14).

La tabla 5 también nos presenta los coeficientes de generalizabilidad, relativos y absolutos, que nos permiten comparar entre 3 y 5 sesiones y entre diferentes planes de estimación en 5 sesiones. La variación en los resultados, en ambas comparaciones, es casi imperceptible o no dan lugar a otra interpretación en los planes de medida 3, 6, 7, 8 y 9.

En el plan 1 detectamos que si reducimos el *universo de generalización* a 10 observadores, $N_o = 10$, la fiabilidad es perfecta. En cierta medida, generalizamos a un número menor de observadores, pero en investigaciones de esta índole es difícil incluso contar con esos 10 observadores entrenados para tal fin. En consecuencia, aumenta la generalización a medida que se reduce el universo de generalización, pero en el caso que comentamos consideramos que la decisión ha sido acertada.

En el plan 2 advertimos un aumento de la generalización, que viene incrementado al pasar de 3 a 5 sesiones (0.90 y 0.94 respectivamente) y un aumento mayor si reducimos el *universo de generalización* de las sesiones a 10, $N_s = 10$ (0.97). En este caso el error de medida ha sido originado por las sesiones, dado que el componente de observadores era prácticamente nulo (ver Tabla 4). La decisión también ha sido acertada, pues difícilmente contaríamos en investigaciones observacionales con un número de sesiones mayor que éste.

Los planes 4, 5, 11 y 14 apenas hacen perceptible una variación en el coeficiente de generalizabilidad. En todos ellos intervienen los individuos como faceta de *generalización*, que en los 3 casos siempre han sido considerados como faceta aleatoria infinita, $N_i = \infty$. En consecuencia, no aumentamos en la precisión de la generalización, pero en contrapartida obtenemos un universo de generalización amplio, ya que los resultados podrán generalizarse a un universo infinito de *individuos* de la población origen de donde seleccionamos las muestras.

En el plan 10, donde sólo las sesiones (S) son la faceta de *instrumentación*, tenemos la mejor evidencia de la estructura que hemos diseñado a través de la tabla 5. Podemos aumentar la precisión de generalización si tenemos en cuenta los registros de dos sesiones más (0.89 y 0.94, respectivamente) y aumentar a 0.97 si se reduce el *universo de generalización* a 10 sesiones, $N_s = 10$. Observamos que los resultados son idénticos a los detectados en el plan 2, lo que pone de manifiesto nuevamente que los observadores no incrementan el error de medida.

Finalmente, los planes 12 y 13 evidencian una disminución del coeficiente de generalizabilidad al pasar de 3 a 5 sesiones y al pasar de estimaciones infinitas a finitas en el caso de observadores y sesiones. Este hecho viene determinado por la escasa variabilidad que aportan las variancias de diferenciación (sesiones en el plan

12 y observadores en el plan 14. Otro hecho destacable es la diferencia existente en los 3 casos entre los coeficientes de generalizabilidad *relativos* y los *absolutos*. Ello viene a demostrar que, aunque los coeficientes son relativamente elevados en 3 sesiones, deben interpretarse con suma precaución. Estas diferencias indicarían que se deberían imponer ciertas condiciones restrictivas en la selección de los niveles de las facetas de generalización. Concretamente, algo que ya hemos propuesto en este trabajo, la clasificación de los individuos según el sexo, que además permitiría reducir el número de niveles de la misma. Al mismo tiempo, la redefinición del *universo de generalización* para investigaciones futuras.

A destacar, por tanto, que los coeficientes de generalizabilidad se refieren al *universo de generalización*. Los mayores beneficios de un análisis de generalización se derivan esencialmente de los resultados obtenidos en esta fase, que permiten llevar a cabo modificaciones en el diseño de medida y elegir un *diseño optimizado* (óptimo en el sentido de que se busca una máxima generalizabilidad dentro de los costos y otras restricciones prácticas o, alternativamente, que se reduzcan los costos mientras se mantega un elevado o aceptable nivel de generalizabilidad). A continuación ilustramos algunas de las posibles modificaciones al Plan de Observación inicial.

DISCUSION Y OPTIMIZACION DE RESULTADOS

Las tablas que se presentan a continuación corresponden a ocho de los planes de medida anteriores, pero modificando sucesivamente el plan de observación original 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. Dado que en algunos de los planes de medida (en el caso de que las categorías sean consideradas *instrumentos de medida*) obtenemos poca precisión relativa de generalización, realizaremos las modificaciones oportunas en cada plan en las facetas consideradas como *instrumentos de medida*. Todas las optimizaciones se han llevado a cabo teniendo en cuenta el Plan de Estimación en el que disponíamos de cinco sesiones, donde $N_i = N_c = N_s = \infty$ y $N_o = 20$.

En el Plan de Medida 2 *IC/OS* (Tabla 6), que nos ofrece un coeficiente de generalizabilidad que podríamos considerar aceptable (0.94), observamos que un aumento del número de sesiones incrementa el coeficiente de generalizabilidad a 0.987 ($\approx$ 0.99) si hubiéramos realizado 20 sesiones de observación. Para conseguir este aumento de precisión en la generalización a sesiones y observadores necesitaríamos realizar 2240 registros en lugar de los 560 iniciales. Luego, la relación costo-beneficio podría considerarse nula y no harían falta modificaciones ni al diseño ni al plan de estimación. Al mismo tiempo, detectamos que el valor de generalización es casi idéntico, tanto si desponemos de 2 como de 3 observadores; en consecuencia, el insignificante error de medida detectado proviene del número de sesiones y no del número de niveles de la faceta observadores.

Tabla 6. Optimización de algunos Planes de Medida en los que las Categorías siempre han sido consideradas como objeto de estudio en el caso de cinco sesiones.

Optimización del Plan de Medida 2 IC/OS

FACETA	OBSERVADOS	ESTIMADOS	OPTIMIZACION DEL PLAN				
I	$n_i = 8$	$N_i = \infty$	8	8	8	8	8
O	$n_o = 2$	$N_o = 20$	2	2	2	3	2
C	$n_c = 7$	$N_c = \infty$	7	7	7	7	7
S	$n_s = 5$	$N_s = \infty$	6	10	15	15	20
$E\rho^2_\delta$	0.94		.957	.974	.982	.983	.987
$E\rho^2_\Delta$	0.94		.952	.971	.980	.981	.985

Optimización del Plan de Medida 10 ICO/S

FACETA	OBSERVADOS	ESTIMADOS	OPTIMIZACION DEL PLAN				
I	$n_i = 8$	$N_i = \infty$	8	8	8	8	8
O	$n_o = 2$	$N_o = 20$	2	2	2	2	2
C	$n_c = 7$	$N_c = \infty$	7	7	7	7	7
S	$n_s = 5$	$N_s = \infty$	6	10	15	20	30
$E\rho^2_\delta$	0.94		.955	.972	.981	.986	.991
$E\rho^2_\Delta$	0.94		.950	.969	.979	.984	.990

Optimización del Plan de Medida 11 COS/I

FACETA	OBSERVADOS	ESTIMADOS	OPTIMIZACION DEL PLAN				
I	$n_i = 8$	$N_i = \infty$	10	15	20	30	50
O	$n_o = 2$	$N_o = 20$	2	2	2	2	2
C	$n_c = 7$	$N_c = \infty$	7	7	7	7	7
S	$n_s = 5$	$N_s = \infty$	5	5	5	5	5
$E\rho^2_\delta$	0.94		.951	.967	.975	.983	.990
$E\rho^2_\Delta$	0.94		.951	.967	.975	.983	.990

Optimización del Plan de Medida 14 OC/IS

FACETA	OBSERVADOS	ESTIMADOS	OPTIMIZACION DEL PLAN				
I	$n_i = 8$	$N_i = \infty$	10	12	15	20	30
O	$n_o = 2$	$N_o = 20$	2	2	2	2	2
C	$n_c = 7$	$N_c = \infty$	7	7	7	7	7
S	$n_s = 5$	$N_s = \infty$	6	8	10	15	20
$E\rho^2_\delta$	0.95		.968	.974	.980	.986	.990
$E\rho^2_\Delta$	0.95		.968	.970	.976	.983	.988

La optimización del Plan de Medida 10 *ICO/S* (Tabla 6), en el que sólo la faceta sesiones (S) ha sido considerada *instrumento de medida*, detectamos resultados similares a los anteriores. Evidentemente, las modificaciones que se realizan en un plan de optimización tan solo afectan a la faceta de *instrumentación*. En consecuencia, aumentar el número de niveles de la faceta sesiones hasta obtener un coeficiente de generalizabilidad relativo cuyo valor sea aproximadamente la *unidad*, conllevaría incrementar a 30 el número de sesiones a realizar en futuras investigaciones. Sin embargo, la relación costo-beneficio también es nula, ya que los costos financieros de la investigación supondrían aumentar a 3360 el número de registros para obtener una *precisión de generalización* de 0.991, cuando obtenemos una precisión de 0.94 en el caso de contar con los 560 registros iniciales. En todo caso, el incremento más importante supondría realizar 10 sesiones y obtener un coeficiente de 0.972, en lugar del 0.94 inicial, aunque supondría *doblar* exactamente el número de registros.

En el caso de que *sólo* los individuos sean considerados como *instrumentación*, plan de medida 11 *COS/I* (Tabla 6), el plan de optimización pone de manifiesto que con $N_i = 10$ la generalización relativa aumenta lentamente (de 0.94 en el diseño original a 0.951 en dicha optimización); si $N_i = 15$, el coeficiente valor 0.967 y el número de registros sería 1050; si $N_i = 20$, 0.975 y 1400 registros; si $N_i = 30$, 0.983 y 2100 registros; y finalmente si $N_i = 50$, 0.99 y 3500 registros. En tal caso, este aumento lento y gradual debe ser valorado por el investigador, quien debe fijar cuál es el grado de generalización que necesita con respecto a la población de individuos de donde extraerá las muestras en las subsecuentes investigaciones. Es decir, ¿podemos generalizar con precisión absoluta a otros individuos extraídos de esa misma población? Es evidente que un pequeño aumento en el número de niveles de la faceta individuos no genera costos excesivos y a cambio se nos ofrece una excelente generalización de los resultados de nuestra investigación.

En el plan de medida 14 *OC/IS* (Tabla 6), las facetas individuos y sesiones han sido consideradas conjuntamente como instrumentos de medida, y por tanto como facetas a generalizar. Como se podrá observar en los diferentes planes de optimización, dado que la variabilidad de la faceta sesiones aporta más error de medida que la faceta individuos, se ha realizado en los cinco casos un incremento mayor en el número de niveles de los individuos que en el de las sesiones. El coeficiente de generalizabilidad tendrá un valor próximo a la *unidad* si dispusiéramos de 30 individuos y 20 sesiones. El diseño original ofrecía unos excelentes resultados, 560 registro y una *precisión de generalización* de 0.95. Sin embargo, obtener la precisión relativa máxima supondría unos costos altísimos, ya que elevaría a 8400 el número de registros a realizar en futuras investigaciones. En consecuencia, la relación costo-beneficio es nula. Sería interesante realizar pequeñas modificaciones al plan original, aumentando el número de niveles en algunas de las facetas de generalización y ello sólo implicaría rediseñar la investigación. Por ejemplo, si $N_i = 10$ y $N_s = 6$, con solo 840 registros tendríamos un valor de precisión en la generalización a individuos y sesiones de 0.968 ($\approx$0.97).

Hemos podido verificar en los cuatro Planes de Medida de la Tabla 6 que los coeficientes eran elevados en el plan original, es decir para las *decisiones* donde interviene la *diferenciación* de las *categorías*. Hemos de formular, sin embargo, una reserva: es verosímil que la estimación de $\sigma^2_{(c)}$ a través de tan sólo cinco categorías sobrevalore la heterogeneidad de éstas y crezca artificialmente la *variancia de diferenciación*. Las otras informaciones buscadas, es decir las que no están ligadas a las *categorías (C)*, y que se presentan en la Tabla 7, son muy poco generalizables; los coeficientes relativos son incluso inferiores a 0.35 en dos planes y los coeficientes absolutos no superan en ningún caso el valor de 0.20. La debilidad del coeficiente de generalizabilidad de los planes 6 y 13 nos revela en particular que el plan de estimación utilizado es suficiente para evaluar con una precisión aceptable el nivel absoluto de los individuos en las categorías. Hará falta aumentar el número de categorías para obtener una media por individuo que sea más fiable. Muestreando un número más grande de categorías en el estudio *G*, estimaremos de forma más precisa la variancia entre categorías (el cuadrado medio y el componente de variancia), ya que la estimación estará basada en un número mayor de grados de libertad. Procediendo de la misma forma en un estudio *D*, disminuiremos las fuentes de error principales de la medida, es decir los errores debidos a las categorías (C) y a sus interacciones con individuos (I), sesiones (S) y observadores (O).

La optimización del plan de medida 6 *ISO/C* (Tabla 7), en ningún caso consigue niveles aceptables de generalización ni de diferenciación. Las cinco categorías utilizadas en la investigación no nos permiten clasificar a los individuos, evaluados por dos observadores, en las sesiones. Aumentando a 40 el número de categorías, *la precisión de generalización* tan sólo alcanza el valor 0.736, siendo necesarios 4000 registros observacionales. Son aconsejables, por tanto, modificaciones al diseño original, aunque los costos de investigación aumenten, dado que las 5 optimizaciones realizadas demuestran que han sido pocas las categorías seleccionadas para estimar con precisión.

El plan de medida 9 *SO/IC* (Tabla 7) ha dispuesto a sesiones (S) y observadores (O) como *objetos de estudio*. Inicialmente, los resultados no son buenos, ahora bien podemos obtener beneficios si aumentamos conjuntamente el número de niveles de las dos facetas de generalización, individuos (I) y categorías (C). Si $N_i = 10$ y $N_c = 10$ la precisión aumenta a 0.856 relizando un total de 1000 observaciones. Aún así estos resultados no dejan de ser un poco artificiales ya que en este mismo caso el coeficiente de generalizabilidad *absoluto* no sobrepasa el valor de 0.262. En todas las demás optimizaciones las diferencias entre coeficientes absolutos y relativos son notables y, por tanto, estas modificaciones deben hacerse con una esmerada selección de individuos y categorías. Necesitaremos probablemente *redefinir* nuestro *universo de generalización* y los individuos y las categorías seleccionadas deberán serlo en función de algún criterio restrictivo que elimine estos importantes errores de medida detectados entre los coeficientes absolutos y relativos. Por tanto, si se realizan modificaciones reales al diseño

original implicará un cambio en las características a definir en la población de individuos. Observamos también que si dispusiéramos de 3000 datos observacionales, $N_i = 15$ y $N_c = 20$ el valor de precisión sería excelente (coeficiente de generalizabilidad *relativo* = 0.933), lo que nos certifica la importancia de la redefinición de las características de la población a la que deseamos generalizar.

En el plan de medida 12 *S/IOC* (Tabla 7) serán sólo las sesiones (S) el *objeto de estudio*. En principio, los resultados son similares al anterior plan de optimización y por tanto no se confirman diferencias significativas entre las diferentes sesiones. De nuevo aparecen diferencias notables entre los coeficientes de generalizabilidad *absolutos* y *relativos*. Ello nos indica la necesaria redefinición del universo de generalización y del aumento del número de categorías, ya que en caso contrario cualquier modificación conllevaría costos altísimos con una relación costo-beneficio, que podríamos considerar nula. Por ejemplo, si $N_i = 10$, $N_o = 3$ y $N_c = 10$, el coeficiente de generalización es de 0.865 (1500 registros) en lugar del coeficiente inicial 0.79 (560 registros). La optimización que mejor regularía los costos significaría aumentar a 15 el número de individuos, a 3 el número de observadores y a 15 el número de categorías, quintuplicando el número total de registros (3375). Aumentar en mayor precisión traería como consecuencia una relación costo-beneficio nula. Por tanto, una precisa diferenciación de las sesiones de observación conllevaría un aumento del número de categorías y probablemente una redefinición de la población de individuos a generalizar. Esta redefinición podría incluir una clasificación de los individuos según el sexo, tal y como hemos apuntado ya en dos ocasiones en el presente trabajo.

Finalmente, el plan de medida 13 *O/ICS* (Tabla 7) tampoco nos augura unas excelentes generalizaciones. Probablemente este plan de medida no tiene interés desde el punto de vista de la optimización, dado que es el plan opuesto al número 1, es decir el que nos ofertaba la fiabilidad interobservadores. Es evidente que si el coeficiente de fiabilidad es alto, tal y como ocurre en nuestro estudio (ver en la Tabla 5 el plan 1 en el que los valores para *3* y *5* sesiones son cercanos a la *unidad*), las diferencias entre observadores no deben ser significativas y por tanto, al no haber variabilidad, es lógico que este plan 13 nos ofrezca resultados de generalización casi nulos. Valga como ejemplo la primera columna optimizada, donde $N_i = 20$ $N_c = 20$ y $N_s = 20$. El valor del coeficiente es 0.2 y el número de datos a registrar sería de 16000. De nuevo se vuelve a poner de manifiesto que el número de niveles de la faceta *categorías (C)* que se han seleccionado para llevar a cabo la investigación han sido insuficientes si se quiere generalizar con cierta precisión. De todas formas, se hace necesaria una valoración de las diferentes relaciones costos-beneficios en estos cuatro planes de la Tabla 7, dado que no es fácil encontrar en este tipo de investigaciones observadores expertos que puedan registrar fiablemente en un sistema con un número mayor de categorías.

Tabla 7. Optimización de algunos Planes de Medida en los que las Categorías siempre han sido consideradas instrumento de medida en el caso de cinco sesiones.

Optimización del Plan de Medida 6 ISO/C

FACETA	OBSERVADOS	ESTIMADOS	OPTIMIZACION DEL PLAN				
I	$n_i = 8$	$N_i = \infty$	8	8	8	8	8
O	$n_o = 2$	$N_o = 20$	2	2	2	3	2
C	$n_c = 7$	$N_c = \infty$	10	15	20	30	40
S	$n_s = 5$	$N_s = \infty$	5	5	5	5	5
$E\rho^2_\delta$	0.32		.411	.511	.582	.677	.736
$E\rho^2_\Delta$	0.15		.203	.277	.338	.434	.561

Optimización del Plan de Medida 9 SO/IC

FACETA	OBSERVADOS	ESTIMADOS	OPTIMIZACION DEL PLAN				
I	$n_i = 8$	$N_i = \infty$	10	12	10	15	15
O	$n_o = 2$	$N_o = 20$	3	3	3	3	3
C	$n_c = 7$	$N_c = \infty$	10	10	12	15	20
S	$n_s = 5$	$N_s = \infty$	5	5	5	5	5
$E\rho^2_\delta$	0.78		.856	.866	.876	.914	.933
$E\rho^2_\Delta$	0.19		.262	.263	.298	.351	.419

Optimización del Plan de Medida 12 S/IOC

FACETA	OBSERVADOS	ESTIMADOS	OPTIMIZACION DEL PLAN				
I	$n_i = 8$	$N_i = \infty$	10	12	10	15	15
O	$n_o = 2$	$N_o = 20$	3	3	3	3	3
C	$n_c = 7$	$N_c = \infty$	10	12	12	15	20
S	$n_s = 5$	$N_s = \infty$	5	5	5	5	5
$E\rho^2_\delta$	0.79		.865	.893	.885	.920	.938
$E\rho^2_\Delta$	0.19		.261	.300	.298	.350	.418

Optimización del Plan de Medida 13 O/ICS

FACETA	OBSERVADOS	ESTIMADOS	OPTIMIZACION DEL PLAN				
I	$n_i = 8$	$N_i = \infty$	20	30	40	50	100
O	$n_o = 2$	$N_o = 20$	2	2	2	2	2
C	$n_c = 7$	$N_c = \infty$	20	30	40	50	100
S	$n_s = 5$	$N_s = \infty$	20	30	40	50	100
$E\rho^2_\delta$	0.24		.207	.338	.442	.521	.727
$E\rho^2_\Delta$	0.00		.000	.000	.000	.000	.001

CONCLUSIONES

La teoría de la generalizabilidad permite distinguir entre un estudio G (*generalizabilidad*), y un estudio D (*decisión* u *optimización*), donde los registros y observaciones nos permiten tomar decisiones o llevar a cabo interpretaciones. Generalmente, primero llevamos a término el estudio de fiabilidad con el fin de determinar los niveles de las facetas de interés que son necesrias para asegurarnos un grado suficiente de precisión de medida en estudios posteriores. En ocasiones, el estudio G y el estudio D se realizan al mismo tiempo. En nuestro caso, el estudio G nos ha permitido valorar si los observadores a través de un sistema de siete categorías pueden evaluar fiablemente a diversos individuos en varia sesiones de observación. Los resultados nos muestran, en general, que el sistema de siete categorías utilizadas en cinco sesiones no son lo suficientemente fiables, es decir sus resultados no pueden generalizarse con precisión al universo de situaciones donde un sistema de 7 categorías ha sido cruzado con 5 sesiones de observación. El investigador debe tener en consideración si puede añadir un mayor número de categorías y/o sesiones a su estudio para conocer los resultados que hallaría en posteriores investigaciones. El estudio D (u optimización del plan) nos permite controlar, en función de la relación costo-beneficio, el número de categoría y/o sesiones necesarias para estimar con precisión y que los resultados puedan ser generalizables.

Concluimos, por tanto, que no serían suficientes las modificaciones al Plan de estimación (ya pusimos de manifiesto el elevado componente de variancia de la faceta categorías [C] y de sus interacciones con individuos [IC] y sesiones [ICS] y su alta contribución al error 61.82, 14.10 y 17.60, respectivamente en el plan de estimación aleatoriamente infinito de 5 sesiones) y se necesitaría una revisión de la faceta *categorías* para plantear con éxito investigaciones futuras. La *fiabilidad interobservadores* es óptima y no exigiría ningún aumento en el número de niveles de esta faceta (O). La estabilidad de las *sesiones* tan sólo implicaría en subsecuentes estudios un aumento del número de niveles de las mismas, mientras que la faceta *categorías* exigiría rediseñar la investigación. Dado que la faceta *individuos* en ningún momento ha supuesto un aumento del error en los 8 planes de optimización considerados, sería conveniente en futuras investigaciones llevar a cabo una anidación de esta faceta, por ejemplo clasificando a los individuos según sexo. De esta forma, restringiríamos los costos en el número de niveles de la faceta *individuos*, determinaríamos una buena precisión y con un número menor de individuos (puesto que estarían clasificados) obtendríamos una información más amplia y podríamos generalizar con mayor precisión. El análisis de generalizabilidad ha servido de estudio piloto para plantear de nuevo una recogida de datos más coherente al diseño de medida.

En consecuencia, el diseño inicial cruzado no nos asegura precisión en los resultados cuando han de generalizarse a un número mayor de muestras y, por tanto, o se eligen nuevos niveles para la faceta *categorías* o se aumenta el número

de las mismas o, finalmente, rediseñamos nuestro estudio en función de la información suministrada por la teoría de la generalizabilidad.

La decisión final correspondería a la optimización de resultados y a la *maximización* de los diferentes coeficientes de generalizabilidad. Para ello, y basándonos en los trabajos de Woodward & Joe (1973), Marcoulides & Goldstein (1990) y Goldstein & Marcoulides (1991), tendríamos que establecer las relaciones costo-beneficio en función de restricciones (presupuestarias y de otro tipo) del diseño de investigación (Arnau, Blanco y Losada, 1990, 1991), con el fin de *maximizar* los coeficientes obtenidos y replanificarlo (si así fuera necesario), ya que nos permitiría determinar el número ideal de registros a realizar en el Plan de Observación inicial. En nuestro caso, no se han maximizado los coeficientes de generalizabilidad, dado que una de las facetas y dos interpretaciones (C, IC, ICS) aportaban mucho error en los componentes de variancia y ello, en principio, suponía costos altísimos en el incremento del número de dicha faceta (C). Si hubiera sido posible, la maximización de los coeficientes en función de los costos reales de la investigación, nos hubiera permitido determinar el número óptimo de observaciones a realizar en nuestro diseño, fijando al mismo tiempo el número óptimo de niveles de cada una de las cuatro facetas.

La ventaja de este tipo de análisis es la capacidad para diseñar estudios D más eficientemente en base a la información aportada por los planes de *estimación* y *medida*. Si tenemos en cuenta la relación nivel de precisión-costos (Johnson & Bell, 1985), es obvio que podremos diseñar un estudio D óptimo.

En definitiva, la teoría de la generalizabilidad organiza la información oponiendo las aplicaciones *exploratorias* a las *confirmatorias*, o bien las aplicaciones *a priori* a las *a posteriori*:

— Una primera dirección del análisis de generalizabilidad se refiere a competencias o dominios todavía no lo suficientemente conocidos. A través del estudio de las fuentes de variancia ligadas a las diferentes facetas, es posible estructurar el conjunto de datos que han sido registrados. Es decir, cuando se han determinado las direcciones principales de la variación, las que son más generalizables, ya podemos saber cuáles son las facetas que deben tenerse en cuenta posteriormente para muestrear esa competencia o dominio. Esta sería utilización heurística de los coeficientes de generalizabilidad que permitiría un análisis *exploratorio* de los datos.

— En otros casos, puede ser que la competencia sea bastante conocida y nos permita constituir dispositivos de observación sistemática. De esta forma, la teoría de la generalizabilidad nos permite obtener conclusiones *confirmatorias* similares a las de las pruebas estadísticas habituales.

— En el espíritu de Cronbach et al. (1972), un análisis *G* constituye normalmente un estudio *a priori*, que sirve para preparar un diseño de investigación a 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 for-

mas, todas las fases de un estudio de generalizabilidad 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.

— 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 diseños de investigación.

Como hemos podido verificar a lo largo de este trabajo, hay una gran diversidad de utilizaciones posibles de la teoría de la generalizabilidad que pueden aplicarse a las diferentes áreas de las ciencias sociales, de la salud o del comportamiento.

REFERENCIAS

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

Anguera Argilaga, M.T. (1985). *Metodología de la observación en las Ciencias Humanas* (3 ed.rev.). Madrid: Cátedra.

Anguera Argilaga, M.T. (1987). Uso de mapas conductuales y cognitivos en evaluación ambiental. En R. Fernández Ballesteros (Ed.), *El ambiente. Análisis psicológico* (pp. 81-102). Madrid: Pirámide.

Anguera Argilaga, M.T. (1988). *Observación en el aula*.Barcelona: Graó.

Anguera Argilaga, M.T. (Ed.) (1991). *Metodología observacional en la investigación psicológica* (vol. 1). Barcelona: P.P.U.

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.

Anguera Argilaga, M.T. y Blanco Villaseñor, A. (1988). Generalizabilidad en la evaluación de mapas conductuales-cognitivos y aplicación de un modelo log-lineal. En J.I. Aragonés y J.A. Corraliza (Eds.), *Comportamiento y medio ambiente: La Psicología Ambiental en España* (pp.673-681). Madrid: Comunidad de Madrid.

Arnau Gras, J., Blanco Villaseñor, A., y Losada López, J. L. (1990, noviembre). *Semejanzas y disimilitudes entre el método multivariable (medidas repetidas) y la teoría de estimación de la precisión (generalizabilidad)*. Comunicación presentada en el VIII Congreso Nacional de Psicología (Mesa Redonda 'Métodos y Técnicas'). Barcelona.

Arnau Gras, J., Blanco Villaseñor, A., y Losada López, J. L. (1991). Estimación de la precisión de un diseño multivariable de medidas repetidas. *Anales de Psicología, 7*, 85-105.

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. (1989). Fiabilidad y generalización de la observación conductual. *Anuario de Psicología, 43*, 5-32.

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.

Blanco Villaseñor, A. y Anguera Argilaga, M.T. (1984, Diciembre). *Avances metodológicos en la evaluación de mapas cognitivos*. Comunicación presentada en el Symposium sobre Actividad Humana y Procesos Cognitivos. Madrid.

Blanco Villaseñor, A., Losada López, J.L. y Anguera Argilaga, M.T. (1991a, Mayo). *Descripción de errores de medida en estudios observacionales de evaluación conductual*. Comunicación presentada en el II Congreso Internacional "LATINI DIES" (Symposium 'La metodología observacional en la evaluación del comportamiento'). Sitges.

Blanco Villaseñor, A., Losada López, J.L. y Anguera Argilaga, M.T. (1991b). Estimación de la precisión en diseños de evaluación ambiental. *Evaluación Psicológica/ Psychological Assessment.*

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., & Allal, L. (1976). The simmetry of generalizability theory: Applications to educational measurement. *Journal of Educational Measurement, 13*, 119-135.

Cardinet, J., Tourneur, Y., & Allal, L. (1981). Extension of generalizability theory and its applications in educational measurement. *Journal of Educational Measurement, 18*, 183-204.

Cardinet, J., & Toruneur, Y. (1985). *Assurer la mesure*. Berne: Peter Lang.

Cardinet, J. (1987). *La construction de tests d'apprentissage selon la théorie de la généralisabilité (Recherches; 87.107). Neuchâtel: Institut romand de recherches et de documentation pédagogiques.*

Chalmers, D.J., & Knight, R.G. (1985). The reliability of ratings of the familiarity of environmental stimuli. A. Generalizability Analysis. *Environment and Behavior, 17* (2), 223-238.

Cronbach, L.J., Rajaratnam, N., & Gleser, G.C. (1963). Theory of gneralizability: a liberalization of reliability theory. *British Journal of Mathematical ans 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., Engelman, L. & Jennrich, R.I. (1990) *BMDP Statistical Software Manual*. Berkeley, Ca.: University of California Press.

Duquesne, F. (1986). Développement sur micro-ordinateur d'un programme pour l'etude de la généralisabilité des données. *Scientia Paedagogica Experimentalis, 23*, 29-36.

Goldstein, Z., & Marcoulides, G.A. (1991). Maximizing the coefficient of generalizability in decision studies. *Educational and Psychological Measurement, spring* (1), 79-88.

Johnson, S., & Bell, J.F. (1985). Evaluating and predicting servey efficiency using generalizability theory. *Journal of Educational Measurement, 22*, 107-119.

Marcoulides, G.A. (1989). The application of generalizability analysis to observational studies. *Quality & Quantity: The International Journal of Methodology, 23*, 115-127.

Marcoulides, G.A. & Goldstein, Z. (1990). The optimization of generalizability studies with resource constraints. *Educational and Psychological Measurement, 50*, 761-768.

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.

Miller, D.C. (1991). *Handbook of research design and social measurement* (5th ed.). Newbury Park, Ca.: Sage Publications.

Mitchell, S.K. (1979). Interobserver Agreement, Reliability, and Generalizability of Data Collected in Observational Studies. *Psychological Bulletin, 86*, 376-390.

Nussbaum, A. (1984). Multivariate Generalizability Theory in Educational Measurement: An empirical study. *Applied Psychological Measurement, 8*, 219-230.

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.

Schroeder, H.W. (1984). Environmental perception rating scales. A case for simple methods of analysis. *Environment and Behavior, 16*, 573-598.

Shadish, W.R., Cook, T.D., & Leviton, L.C. (1991). *Foundations of program evaluation.* Newbury Park, Ca.: Sage Publications.

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.(1991). *A Primer on Generalizability Theory*. Newbury Park, Ca.: Sage Publications.

Shavelson, R.J., Webb N.M., & Rowley, G.L. (1989). Generalizability Theory. *American Psychologist, 44*, 922-932.

Smith, P.L., & Teeter, P.A. (March, 1982). *The Use of Generalizability Theory with Behavioral Observation*. Paper presented at the Annual Meeting the American Educational Research Association. New York.

Streiner, D.L., & Norman, G.R. (1989). *Health Measurement Scales. A Practical Guide to their Development and Use*. New York: Oxford University Press.

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*.

Woodward, J.A., & Joe, G.W. (1973). Maximizing the coefficient of generalizability in multifacet decision studies. *Psychometrika, 38*, 173-181.