Psicothema
2000. Vol. 12, Supl. nº 2, pp. 122-126

ISSN 0214 - 9915 CODEN PSOTEG

Validez en evaluación de programas: una comparación de técnicas de análisis basadas en modelos estructurales, teoría de la generalizabilidad y modelos multinivel

Salvador Chacón Moscoso, José Antonio Pérez-Gil y Fco. Pablo Holgado Tello
Universidad de Sevilla

En el presente trabajo, a partir de un análisis conceptual de la validez en evaluación de programas, se presenta un estudio comparativo de las implicaciones e interacciones que tienen una serie de técnicas de análisis para potenciar el logro de la validez en evaluación. Concretamente se aplica la teoría de la generalizabilida, los modelos de ecuaciones estructurales y los modelos multinivel sobre los datos recogidos en la evaluación del programa de formación continua del personal de Administración y Servicios de la Universidad de Sevilla.

Validity in program evaluation: A comparative approach. Begining from a conceptual analysis of program evalution validity, this paper introduces a comparative study of implications and interrelationships of different statistical techniques in order to enhance validity in program evaluation. Generalizability theory, hierarquical linear analysis and structural equation modeling are applied to an evalution of a training program of administrative staff in University of Sevilla.

Las condiciones inestables del contexto en el que se realiza la evaluación de la mayor parte de los programas de intervención hacen que sea imposible plantear estructuras estándares de diseños de evaluación. Esta situación provoca que la evaluación se ha de configurar de acuerdo con las necesidades, características y elementos moduladores del programa concreto a evaluar. A pesar de la gran cantidad de posible desarrollos evaluativos con los que nos podemos encontrar en un ámbito de intervención tan variado, en cualquier proceso evaluativo se persigue obtener una información de calidad sobre un determinado programa de intervención para alcanzar los fines perseguidos. En nuestro caso consideramos información de calidad a la información válida obtenida desde los criterios de la metodología científica, y es por ello que consideremos el concepto global de validez científica como el pilar a partir del cual justificar el desarrollo de cualquier proceso evaluativo.

Existen un gran número de trabajos en los que se tratan las distintas tipologías de validez en investigación (sólo considerando algunas aportaciones recientes en castellano podrían citarse entre otros: Arnau, Anguera y Gómez, 1990; Ato, 1991; Vallejo, 1991; García Jiménez, 1992; Anguera, 1995; Martínez-Arias 1995); no obstante, en tanto que el objeto de este epígrafe no es presentar un barrido de los distintos desarrollos sobre el tema, sino más bien hacer una introducción al concepto de validez, introducimos directamente la nueva conceptualización del término asumida por Shadish, Cook y Campbell (en preparación). Consideramos que este planteamiento recoge los último desarrollos sobre el concepto de validez aplicado a la evaluación de programas de intervención.

Shadish, Cook y Campbell (en preparación), utilizando la terminología de Cronbach, plantean que la *validez de conclusión estadística* se refiere al estudio de la correlación entre «t» (tratamiento) y «o» (resultado «outcome») en los «utost$_i$[1]». Es decir, si el tratamiento (en este caso programa de intervención), tal y como se ha implementado, correlaciona significativamente con los resultados, tal y como se han medido.

Por su parte la *validez interna* se refiere a si la relación entre t (tratamiento-programa) y «o» (resultado) es causal[2], es decir si los resultados observados se hubiesen dado sin estar presente el programa.

La *validez de constructo* implicaría realizar inferencias sobre la población, o constructos delimitados a partir de la muestra particular de «utost$_i$» utilizada en el programa de intervención; es decir las transferencias de los «utost$_i$» a los «UTOST$_i$».

En último término la *validez externa* se conceptualiza como la posibilidad de generalizar la relación causal estudiada en los «utost$_i$» a otras poblaciones diferentes a las usadas en la intervención «*UTOST$_i$».

Esta tipología reconceptualiza los términos de validez de constructo y validez externa de Cook y Campbell (1979). Es decir, previamente la validez de constructo estaba limitada a las causas y efectos (''t'' y ''o'') y la validez externa se refería sólo a la generalización de las unidades de estudio («u»), contextos (settings«s») y momentos («t$_i$»). Con la nueva conceptualización, tanto la validez de constructo como la externa se aplican a los cinco elementos («u,t,o,s,t$_i$»), aunque de un modo distinto. Ahora la validez de constructo pretende la generalización a la población delimitada en la intervención a través de los elementos muestreados, y la validez

Correspondencia: Salvador Chacón Moscoso
Facultad de Psicología
Universidad de Sevilla
41005 Sevilla (Spain)
E-mail: schacon@psicoexp.us.es

externa persigue la generalización a poblaciones diferenciales de las utilizadas en la intervención.

Con esta nueva conceptualización de los tipos de validez los autores han pretendido dar el mismo peso a la validez interna que a la externa, evitando el sesgo de la versión anterior (Cook y Campbell, 1979) hacia la validez interna. Para ello plantean como la validez externa se basa en los mismos principios que justifican la generalización en la validez de constructo. La validez de constructo hace referencia a la extrapolación a constructos a través de las operaciones de muestreo realizadas en las intervenciones implementadas, y por su parte la validez externa se plantea la extrapolación a constructos diferentes a partir de esas mismas operaciones realizadas en las intervenciones. Por tanto, en ambos casos se están analizando los mismos principios que sustentan el estudio de los constructos.

A modo de resumen, en la figura 1, derivada del esquema presente en Cook, Shadish y Peracchio (1990), aparece la evolución comparativa de la conceptualización del término de validez en los últimos 25 años.

Elementos definitorios del «macroconcepto» de validez en evaluación de programas

Tratando de avanzar más en la integración del concepto de validez presentaremos los elementos definitorios de lo que podríamos denominar el «macroconcepto» de validez intentando fundamentalmente no relegar a un segundo término los estudios no experimentales, cuyo interés no ha de ser exclusiva y necesariamente relacional causal. Con ello pretendemos aportar los elementos que sirvan de punto de referencia para el logro de la validez en el desarrollo de un programa de evaluación, desde el momento inicial en que se esta diseñando hasta la confección del informe final de evaluación.

Un programa de intervención objeto de evaluación puede existir al margen del evaluador, pero no puede ser evaluado al margen de éste; de ahí que referirnos a las características de los programas objetos de evaluación equivale a hacerlo de lo que los evaluadores consideran como tales. Lo que ocurre en la realidad no es descriptible sino según los conceptos determinados, como los de los médicos, enfermeros, sociólogos, pedagogos, economistas, educadores sociales, psicólogos, etc, de ahí que en una misma situación puedan existir diferentes interpretaciones, tantas como conceptos referentes utilicemos.

Durante la exposición de los distintos tipos de validez presentados en la introducción de forma implícita se planteaba una dicotomización entre realidad objeto de estudio y la concepción que de ella se pueda tener (dominios sustantivo y conceptual del proceso de investigación, en la terminología de Brinberg y Kidder, 1982). En realidad en el momento de desarrollar las evaluaciones se convierten en una unidad, ya que los datos que estamos tomando de la «realidad» dependen del referente (programa de intervención en este caso) que hayamos delimitado.

En términos globales el «macroconcepto» de validez se refiere a la medida en que se de una correspondencia de semejanza entre las características del concepto planteado y los datos obtenidos sobre dicho concepto. Es el criterio que Schmitt (1995) denomina de correspondencia teórica. No estamos definiendo exclusivamente el planteamiento de la validez como correspondencia «concepto-dato», ya que en primer lugar, los mismos datos con lo que se valida un referente conceptual tienen un anclaje conceptual o teórico, por lo que no pueden ser considerados como un elemento de comparación en sí mismo libre de planteamiento teóricos (Kuhn, 1962), y en segundo lugar, la inestabilidad de las condiciones sociales en las que se evalúa un programa hace necesario que se tenga en cuenta si los datos obtenidos se relacionan con las teorías de programación previamente existentes, y si dichos resultados a su vez

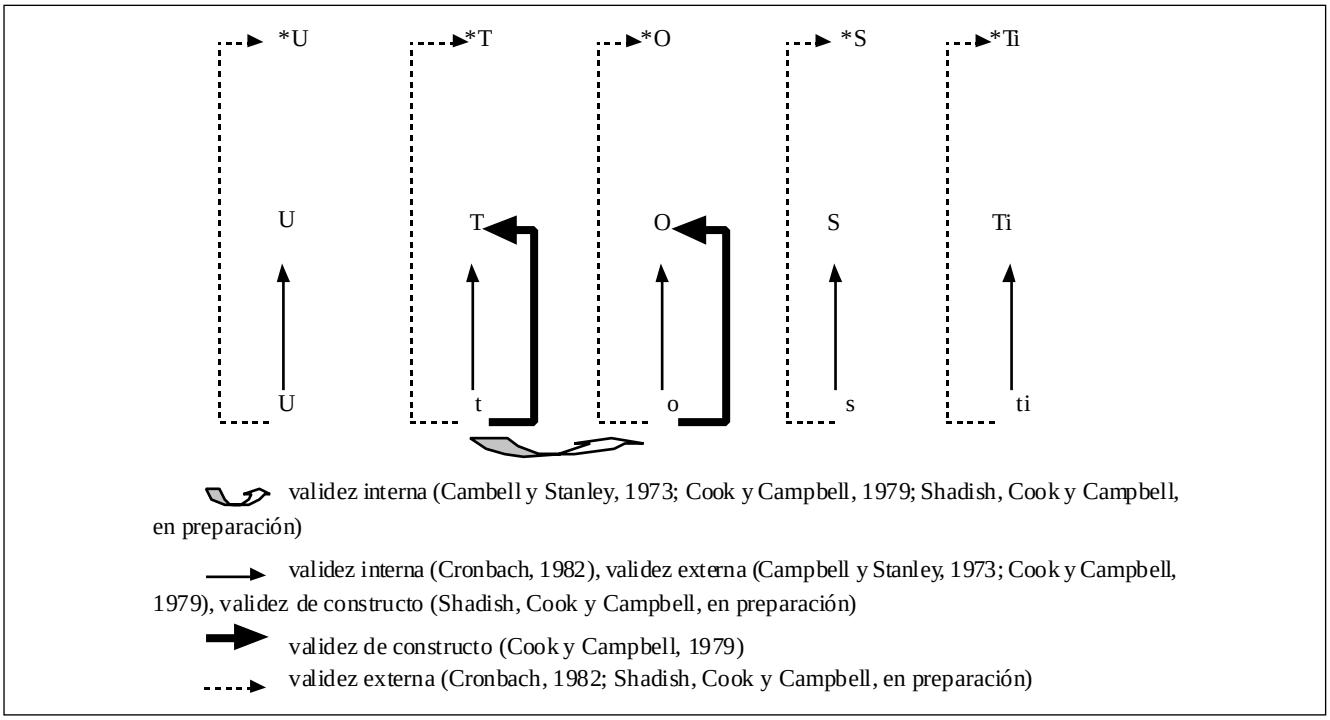

Figura 1. *Evolución comparativa de la conceptualización del término de validez*

son útiles para los implicados en el programa. Por estos motivos, en este trabajo se plantea que la validez de un concepto (programa evaluado) está relacionada, además de con los datos utilizados, con los estudios previos existentes y la utilidad que dicho concepto tenga. Los resultados evaluativos serán validos si son coherentes con un conjunto de conceptos previamente establecidos y si resultan útiles para los implicados en el programa de intervención. A lo largo del proceso evaluativo estos tres criterios (correspondencia, coherencia y utilidad) se han de ir combinando para potenciar el logro de la validez pretendido.

La línea argumental desarrollada nos permite entender la frecuente recomendación en ciencia de que para obtener datos válidos ha de delimitarse adecuadamente lo que se desea estudiar. Es decir, tal y como lo hemos planteado, que las evidencias obtenidas sean representativas y no presenten confusiones con respecto a los elementos definitorios del concepto de referencia. No obstante, dada la complejidad de la mayoría de los programas, es lógico entender que no se pueda dar seguridad absoluta a la validez de un programa de evaluación por la propia definición relativa del término; y porque tal y como plantean autores como Cronbach (1982, 1989), nos estamos refiriendo a la emisión de un juicio subjetivo y por tanto dependiente del criterio referente que se use para realizar dicha valoración. De igual forma Messick (1989) concibe la validez como un juicio evaluativo integrador del grado en que la evidencia empírica y los modelos teóricos apoyan la adecuación de inferencias basadas en determinadas puntuaciones u otros modelos de valoración.

Llegados a este punto nos encontramos ante un bucle cerrado, es decir empezamos enfatizando la necesidad de obtener información válida para acabar matizando que esas garantías de validez no pueden establecerse a no ser que se plantee con nitidez el referente conceptual; a su vez, la validez de dicho referente vendrá dada por su correspondencia con las evidencias obtenidas, teorías previas y la utilidad que presente dicho referente para los distintos participantes en el programa. Este proceso nos lleva a defender el modelo de evaluación en que se enfatiza cómo la evaluación ha de estar guiada desde la teoría (Chen y Rossi, 1987; Lipsey y Pollard, 1989; Chen, 1990). Todo ello se relaciona directamente con la conclusión en la que se plantea cómo la validez de constructo es la principal de los tipos de validez, en tanto que «la validez de constructo es el concepto unificador que integra las consideraciones de contenido y criterio en un marco común para probar hipótesis acerca de relaciones teóricamente relevantes» (Messick, 1980; p.1015), en este mismo sentido (Cronbach, 1984; p.126) señala que «la meta final de la validación es la explicación y comprensión y por tanto ésto nos lleva a considerar que toda validación es validación de constructo». En estos desarrollos se apoya el último trabajo de Shadish, Cook y Campbell (en preparación) en el que la validez de constructo es la piedra angular a partir de la cual se desarrollan los distintos tipos de validez propuestos.

Justificación del uso de algunas técnicas de análisis para estudiar la validez en evaluación de programas

Todo programa de intervención, de acuerdo con unos planteamientos previos (que pueden o no responder a un modelo teórico contrastado), implementa unas acciones en el medio para lograr una serie de objetivos. El contexto inestable en el que se diseñan y ejecutan las evaluaciones hacen que juegue un papel muy importante el modelo teórico o población referente a partir de la cual se ha diseñado el programa de intervención (siguiendo el acrónimo «UTOSTi» de Shadish, Cook y Campbell, en preparación). Esta circunstancia ha hecho que planteemos un concepto de validez unitario referido al grado en que los datos recogidos y el modelo teórico subyacente al programa apoyan las actuaciones e interpretaciones basadas en la información obtenida.

En el caso de la evaluación hemos enfatizado que la validez de la información está relacionada con el uso que los implicados en el proceso de «intervención-evaluación» realicen con los datos obtenidos. Este planteamiento está ligado al concepto de validez consecuencial, y por tanto asume que las cuestiones de valor y consecuencias de las inferencias y acciones a ejecutar forman parte del proceso de validación.

Toda evaluación implica la emisión de un juicio de valor que puede afectar no sólo al programa evaluado en cuestión sino a otros futuros; de ahí la importancia de la generalización de la validez. Es decir, aplicar la evidencia de validez obtenida en un proceso evaluativo en una o más situaciones a otras situaciones de intervención.

Uno de los mayores problemas que se plantean en la evaluación de programas es que al estar inserta en un contexto continuamente cambiante implica que no se suelen dar las condiciones para la aplicación de un diseño de intervención/evaluación estándar con unos instrumentos de registro predeterminados (Anguera, 1994). Por este motivo, además de la elaboración e implementación de diseños particulares adaptados a casuísticas concretas, se ha de recurrir frecuentemente a instrumentos de elaboración propia. Dichos instrumentos suelen ser una fiel traducción del proceso evaluativo desarrollado en tanto suponen el medio a través del cual registrar las conductas a estudiar, previamente definidas en el programa de intervención, de acuerdo con una serie de condicionantes de orden teórico y de variables de contexto implicadas en el proceso.

Por todo ello juega un papel importante la fiabilidad-«calidad» de los datos recogidos, ya que éstos en gran medida nos informarán de forma indirecta de la validez del proceso desarrollado. Es decir, la obtención de unos datos fiables puede ser indicativo de su validez, en tanto su falta de fiabilidad nos informaría inequívocamente de ausencia de ésta. De esta forma, la posibilidad de analizar la influencia de distintas fuentes de variación sobre la fiabilidad obtenida nos permite conocer las variables moderadoras del diseño evaluativo ejecutado, lógicamente teniéndose presente que durante todo el proceso de evaluación se han desarrollado procedimientos que potencien en la mayor medida posible la validez del dato final obtenido.

En el marco del estudio de la validez evaluativa consideramos que la teoría de la generalizabilidad es muy útil. Ofrece un planteamiento globalizador sobre la temática al posibilitar el estudio de la incidencia de distintas variables en el diseño planteado. Este procedimiento permite analizar la «heterogeneidad de las irrelevancias», en el sentido de estudiar de una manera flexible la incidencia (variabilidad explicada) por las distintas variables consideradas en el proceso evaluativo (Shavelson y Webb, 1991).

Las principales características que favorecen la aplicabilidad de la teoría G. en evaluación de programas podrían resumir de forma esquemática en:

- integrar distintas fuentes de variación en una estructura global de diseño operativizada en índices de fiabilidad.
- reconocimiento explicito de distintas fuentes de error.
- posibilidad de analizar distintos objetos de medida desde el mismo diseño.

– estudio sistemático de distintas fuentes de sesgo
– obtención de coeficientes de generalizabilidad «claros» al obligarse a explicitar el universo de generalización.

En definitiva, el objetivo de la teoría G es desglosar en la mayor medida posible, en cualquier tipo de medición, la variabilidad «real» en la/s variables objeto de interés de la variabilidad del error.

En el cuadro 1, a modo ilustrativo, se presentan algunos análisis, aplicando la teoría G, a los datos obtenidos en la evaluación de los programas de formación del PAS, de la Universidad de Sevilla. En estos análisis se puede apreciar qué variables son las más relevantes (mayor variabilidad explicada), en términos de coeficientes de generalizabilidad, dentro de todas aquellas consideradas en los posibles distintos diseños evaluativos planteados.

Al mismo tiempo que se analiza la distribución diferencial de las distintas fuentes de variación en nuestro diseño evaluativo desde la teoría G, pueden utilizarse otras técnicas de análisis complementarias para obtener una mayor información, y por tanto poder tener más elementos de juicio a la hora de emitir valoraciones sobre el programa objeto de evaluación.

En este sentido, puede hacerse uso de los modelos lineales jerárquicos, en los que es posible analizar las aportaciones concretas de variables (parámetros de regresión obtenidos en los modelos jerárquicos) en relación con los coeficientes de generalizabilidad obtenidos desde distintos modelos de medida (Kreft y Leeuw, 1998), al mismo tiempo que permite la dicotomización de predictores de parámetros de cambio (Bryk y Raudenbush, 1992).

Por otra parte, se pueden utilizar los modelos de ecuaciones estructurales, como un modo indirecto del estudio de validez, en el sentido de intentar operativizar los procesos y relaciones entre los distintos factores que están incidiendo en la consecución final de los resultados. De esta forma, se puede intentar representar las hipótesis del programa de intervención a evaluar sobre el modelo estructural de relaciones entre las variables latentes y modelos de medida (de esas latentes) respecto a indicadores observables (Shadish, Cook y Campbell, en preparación). De igual forma, se puede plantear una complementariedad con los análisis anteriores al estudiarse la distinta dimensionalidad de los datos dependiendo del nivel jerárquico de la variable considerada (Hox, 1995).

En el cuadro 2, presentamos una aplicación de los modelos multinivel y de análisis factorial sobre los datos analizados previamente desde la teoría G. En estos análisis se muestra la incidencia concreta de cada una de las variables consideradas en el análisis

Cuadro 1
Análisis desde la Teoría de la Generalizabilidad

Niveles de facetas:
Tipos Cursos: 2
Niveles Cursos: 2
Cursos: 32
Items:11
Sujetos: 960

RESUMEN
Diseños de medida

Plan de medida	σ^2_τ	σ^2_δ	σ^2_Δ	$E\rho^2_\delta$	$E\rho^2_\Delta$
N/TCSI	.05823	.010	.036	.848	.619
	.06691	.002	.010	.950	.764
T/NCSI	.03351	.010	.048	.763	.410
	.04218	.002	.010	.924	.671
I/TNCS	.09603	.005	.055	.955	.635
	.09457	.005	.005	.954	.950
C/SI	.06818	.020	.029	.772	.703
	–	–	–	–	–
I/CS	.09468	.004	.007	.955	.933
	–	–	–	–	–

Cuadro 2
Análisis multinivel y análisis factorial

Análisis multinivel multivariado
(influencia del tipo y nivel de curso en cada una de las vvdd)

$$y_{1jk} a\ y_{11jk} \sim N(XB,\Omega)$$
$$y_{1jk} = \beta_{0jk}\, x_{0ijk} - .308x_{11ijk} - .454x_{22ijk};\ \beta_{0jk} = 4.473 + v_{0k} + u_{0jk}$$
$$y_{2jk} = \beta_{1jk}\, x_{1ijk} - .331x_{12ijk} - .435x_{23ijk};\ \beta_{1jk} = 4.505 + v_{1k} + u_{1jk}$$
$$y_{3jk} = \beta_{2jk} x_{2ijk} - .262x_{13ijk} - .242x_{24ijk};\ \beta_{2jk} = 3.625 + v_{2k} + u_{2jk}$$
$$y_{4jk} = \beta_{3jk} x_{3ijk} - .446x_{14ijk} - .454x_{25ijk};\ \beta_{3jk} = 4.667 + v_{3k} + u_{3jk}$$
$$y_{5jk} = \beta_{4jk} x_{4ijk} - .246x_{15ijk} - .312x_{26ijk};\ \beta_{4jk} = 4.721 + v_{4k} + u_{4jk}$$
$$y_{6jk} = \beta_{5jk} x_{5ijk} - .102x_{16ijk} - .373x_{27ijk};\ \beta_{5jk} = 4.134 + v_{5k} + u_{5jk}$$
$$y_{7jk} = \beta_{6jk} x_{6ijk} - .290x_{17ijk} - .415x_{28ijk};\ \beta_{6jk} = 4.611 + v_{6k} + u_{6jk}$$
$$y_{8jk} = \beta_{7jk} x_{7ijk} - .154x_{18ijk} - .554x_{29ijk};\ \beta_{7jk} = 4.333 + v_{7k} + u_{7jk}$$
$$y_{9jk} = \beta_{8jk} x_{8ijk} - .410x_{19ijk} - .156x_{30ijk};\ \beta_{8jk} = 4.597 + v_{8k} + u_{8jk}$$
$$y_{10jk} = \beta_{9jk} x_{9ijk} - .417x_{20ijk} - .396x_{31ijk};\ \beta_{9jk} = 4.802 + v_{9k} + u_{9jk}$$
$$y_{11jk} = \beta_{10jk}\, x_{10ijk} - .231x_{21ijk} - .277x_{32ijk};\ \beta_{10jk} = 4.703 + v_{10k} + u_{10jk}$$
$$-.2 * log(like) = 24366,800$$

Análisis factorial global

Varianza total explicada

	Autovalores iniciales		
Componente	**Total**	**% de la varianza**	**% acumulado**
1	3.878	37.370	37.370
2	1.243	11.975	49.345
3	1.135	10.933	60.277

Matriz de pesos

	Componente		
	1	**2**	**3**
Ítem 1		-.642	
Ítem 2		-.588	
Ítem 3		-.966	
Ítem 4	.542	-.316	
Ítem 5	.443		
Ítem 6			-.953
Ítem 7	.564		
Ítem 8	.691		
Ítem 9	.755		
Ítem 10	.570		
Ítem 11	.554		

Análisis factorial Tipo de curso= 2

Varianza total explicada

	Autovalores iniciales		
Componente	**Total**	**% de la varianza**	**% acumulado**
1	4.208	39.578	39.578
2	1.245	11.712	51.291
3	1.164	10.952	62.242

Matriz de pesos

	Componente		
	1	**2**	**3**
Ítem 1		-.568	
Ítem 2		-.681	
Ítem 3		-.991	
Ítem 4	.307	-.468	
Ítem 5	.516		
Ítem 6	.942		
Ítem 7	.469	-.321	
Ítem 8			-.866
Ítem 9			-.766
Ítem 10	.517		-.306
Ítem 11	.634		

global de variabilidad, obtenida desde los coeficientes de generalizabilidad, operativizándolos en los pesos concretos de los parámetors de regresión obtenidos desde el análisis multinivel. De igual forma se muestra cómo se distribuyen las distintas saturaciones de variables considerando un tipo u otro de nivel jeráquico de las variables consideradas en el análisis factorial. Estos análisis permiten estudiar la distribución diferencial de las variaciones en uno y otro caso.

En síntesis con este planteamiento queremos enfatizar que el objetivo final es disponer de la información con el mayor grado de validez posible, para a posteriori intentar utilizar distintos análisis complementarios que puedan dar la mayor riqueza de información a la hora de valorar el programa.

Notas

1 Los autores han diferenciado la dimensión tiempo («t_i») del contexto («s») planteando la necesidad de su estudio aislado en el análisis de la validez.

2 Sin considerar la relación causal entre «t» y «o», este sistema de códigos también plantea la posibilidad de aplicarse a otros estudios no necesariamente causales, como algunos estudios observacionales o de encuesta, en los que se pueden encontrar de igual forma elementos referidos a las unidades de estudio, intervenciones, resultados, contextos y momentos de estudio, sin tener que plantear necesariamente relaciones causa-efecto.

Referencias

Anguera, M.T. (1994). El psicólogo en la valoración de programas de intervención. Documento no publicado de la conferencia impartida en la III Semana Psicológica *«La professió del psicòleg»*. Universitat Rovira i Virgili. Tarragona.

Anguera, M.T. (1995). Diseños. En R. Fernández Ballesteros (Ed.) *Evaluación de programas. Una guía práctica en ámbitos sociales, educativos y de salud* (pp.149-172). Madrid: Síntesis.

Arnau, J.; Anguera, M.T. y Gómez, J. (1990). *Metodología de la investigación en ciencias del comportamiento.* Murcia: Servicio de publicaciones de la Universidad de Murcia.

Ato, M. (1991). *Investigación en ciencias del comportamiento. I: Fundamentos.* Barcelona: P.P.U.

Brinberg, D. y Kidder, H. (Eds.) (1982). *Forms of validity in research.* San Francisco: Jossey-Bass.

Bryk, A. y Raudenbush, S. (1992). *Hierarquical linear models.* Londres: Sage.

Campbell, D. y Stanley, J. (1973). *Diseños experimentales y cuasiexperimentales en investigación educativa.* Argentina: Amorrortu Editores.

Chen, H. y Rossi, P.H. (1987). The Theory-driven approach to validity. *Evaluation and Programm Planning, 10,* 95-103.

Chen, H. (1990). *Theory-driven evaluations.* London: Sage.

Cook, T. y Campbell, D. (1979). *Quasi-experimentation. Design and analysis issues for field settings*. Boston: Houghton Mifflin.

Cronbach, L. (1982). *Designing evaluations of educational and social programs*. San Francisco: Jossey-Bass.

Cronbach, L. (1984). *Essentials of psychological testing*. 4ª Edic. Nueva York: Harper & Row.

Cronbach, L. (1989). Construct Validation after thirty years. En Linn, R.L. (ed.). *Intelligence: Measurement, theory and Public policy. Urban and Chicago*. University of Illinois Press.

García Jiménez, M.V. (1992). *El método experimental en la investigación psicológica*. Barcelona: P.P.U.

Hox, J. (1995). Applied multilevel analysis. Amsterdam: TT-Publikaties.

Khun, T.S. (1962). *The structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Kreft, I y Levuw, J. (1998). *Introducing multilevel modeling*. Londres: Sage.

Lipsey, M.W. y Pollard, J.A. (1989). Driving toward theory in program evaluation: more models to choose from. *Evaluation and Program Planning, 12,* 317-328.

Martínez Arias, R.M. (1995). *Psicometría: teoría de los tests psicológicos y educativos*. Madrid: Síntesis.

Messick, S. (1989). Validity. En Linn, R.E (ed) Educational measurement. National Council of measurement in education. Series on Higher Education Oryx Press (pp. 13-102).

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35,* 1012-1027.

Schmitt, F.F.(1995). *Truth: A primer.* Boulder, Colorado: Westview Press.

Shadish, W.R.; Cook, T.D. y Campbell, D.T. (en preparación). *Experimental and quasi-experimental designs for generalized causal inference.* Boston: Houghton-Mifflin.

Shavelson, R.J. y Webb, N.M. (1991). *Generalizability theory. A primer.* Londres: Sage.

Vallejo, G. (1991). La validez de la investigación en el ámbito experimental. En J. Pascual, M.T. Anguera, G. Vallejo y F. Salvador. *Psicología experimental* (pp.41-77). Valencia: Nau Llibres.