El II Congreso Mundial Vasco, patrocinado por el Gobierno Vasco y organizado con la colaboración de las más prestigiosas instituciones académicas y culturales del País Vasco, se ha celebrado entre los meses de agosto y diciembre de 1987. Se trata de un encuentro multidisciplinar que ha agrupado a 34 Congresos y «Workshops», cubriendo diferentes áreas y campos de conocimiento.

El II Congreso Mundial Vasco ha contemplado un doble objetivo. En primer lugar, tratar temas de actualidad desde una perspectiva científica y universal, en el intento de encontrar soluciones a problemas de interés social, y en segundo lugar, ofrecer un marco adecuado para el fortalecimiento de vínculos entre el País Vasco y la Comunidad cultural y científica internacional, al mismo tiempo que ha servido de oportuna plataforma para el intercambio de ideas, métodos y recomendaciones en el campo de la Ciencia, la Técnica, la Cultura, etc.

En este marco, el Congreso de Educación, celebrado en Bilbao, entre el 13 y el 17 de octubre de 1987, ha reunido las más modernas aportaciones de los más prestigiosos especialistas en seis áreas que comprenden la temática educativa de nuestro momento: planificación de la educación y mercado de trabajo, la gestión educativa ante la innovación y el cambio, temas actuales sobre psicopedagogía y didáctica, perspectivas y problemas de la función docente, tecnología y educación, y aspectos metodológicos de la investigación educativa.

Congreso de Educación / 6

II CONGRESO MUNDIAL VASCO

narcea

Aspectos metodológicos de la investigación educativa

6

Aspectos metodológicos de la investigación educativa

IÑAKI DENDALUCE (Coord.)

Margarita Bartolomé
Thomas Cook
Törsten Husén
Robert Linn
Joan Mateo
Mario de Miguel
Arturo de la Orden
Laura Peracchio
Philippe Pilibossian
Javier Tejedor
...

II CONGRESO MUNDIAL VASCO

narcea

LIMITACIONES

En el momento de entrega del presente trabajo no ha sido posible ofrecer resultados del estudio debido a problemas encontrados al aplicar el programa LOGIST. Dicho programa contiene una serie de subrutinas que limitan el uso del programa a un IBM 360 al cual no tenemos acceso. Puestos en contacto con Wirgensky, se nos ha notificado que se están haciendo intentos por adaptarlo para su uso más generalizado.

Actualmente se pretende aplicar el programa ASCAL. Habrá que estudiar sus propiedades y limitaciones respecto al LOGIST, por lo que quizá habrá que limitar y ajustar nuestro estudio a las posibilidades que ofrezca dicho programa.

REFERENCIAS BIBLIOGRAFICAS

BERK, R. A. (1982): *Handbook of Methods for detecting test bias*. The Johns Hopkins University Press, Baltimore.

COLE, N. S. (1981): «Bias in testing». *American Psychologist, 10*, 1067-1077.

HAMBLETON, R. K. y SWAMINATHAN, H. (1985): *Item response theory: Principle and Applications*. Kluwer, Nijhofthus, Dordiecht. Boston.

IRONSON, G. H. y SUVKOVIAK, M. J. (1979): «A comparisons of several methods of assessing item bias». *Journal of Educational Measurement, 16*, 209-225.

LORD, F. M. (1980): *Applications of item response theory to practical testing problems*. Addison-Wesley, Reading, Mass.

WINGERSKY, M. S.; BARTON, M. A. y LORD, F. M. (1982): *LOGIST user'guide*. Educational testing Service, Princenton, Nueva Jersey.

Begoña ORDEÑANA GARCIA. Universidad del País Vasco. Facultad de Filosofía y Ciencias de la Educación. Departamento de Métodos de Investigación y Diagnóstico en Educación. Alto de Zorroaga, s/n. 20014 San Sebastián (España).

1987

3

Aplicación de los modelos Log-Lineales para el análisis de elementos en pruebas de referencia criterial (TRC)

J. M. Jornet y J. M. Suárez

Las características propias de los *ítems* de Tests de Referencia Criterial, orientados a medir niveles mínimos de competencia conllevan que la mayoría de indicadores basados en el Modelo Psicométrico Clásico (Lord y Novick, 1968; Lord, 1980) carezcan de sentido y, en todo caso, deban ser reinterpretados en su uso para medir la Sensitividad Instruccional. Este concepto, definido originalmente por Kosekoff y Klein (1974), se presenta actualmente como la característica central de los *ítems* de los TRC (Haladyna y Roid, 1981; Roid y Haladyna, 1982; Berk, 1984; Jornet, 1987) y se entiende como la capacidad que tiene el *ítem* para detectar los cambios debidos a un programa educativo. Como tal alude pues a la validez interna de la prueba. No vamos a entrar aquí, puesto que lo hemos hecho en otra parte (Jornet y Suárez, 1987), en la cuestión referida a la clarificación del concepto en base al diseño y sus diversas alternativas y consecuencias. Más bien, pretendemos revisar sintéticamente las alternativas que se han propuesto para cuantificar este parámetro y a continuación sugerir una alternativa, valorando al mismo tiempo la aportación en términos comparativos con el resto de indicadores y con las exigencias que entendemos plantea la medición del concepto.

La operacionalización del concepto de sensitividad instruccional requiere, junto a la revisión lógica *a priori*, un análisis empírico *a posteriori* y el establecimiento de un *feed-back* lógico-empírico congruente con el conjunto de las etapas de construcción de un TRC. Así, por lo que se refiere a los procedimientos de cómputo propuestos como indicadores de sensitividad instruccional, estos han sido extremadamente variados y, en general, insatisfactorios para aportar información consistente respecto al parámetro que pretenden medir. Esto es así dado que los datos disponibles para tal cometido (normalmente, pretest-postest sobre un *ítem* dicotómico) hacen difícil su análisis, por violar la mayor parte de las veces supuestos básicos de muchas aproximaciones de contraste de hipótesis.

Los métodos más relevantes se recogen, por categorías metodológicas, en la tabla 1, donde se aprecia la dispersión que acabamos de apuntar. La exposición valorada de los procedimientos se ha realizado en otro trabajo (Jornet y Suárez, 1987) por lo que aquí entraremos directamente en la descripción del acercamiento que nos ocupa.

En el marco que acabamos de referir, hemos propuesto un procedimiento sustentado en los modelos log-lineales que, entendemos, se ajustan mejor a las características de la información disponible y proporcionan unos indicadores precisos

sobre un conjunto de parámetros de interés. La primera ventaja de este tipo de modelos reside en que se desenvuelven con variables categoriales, lo que se adecúa muy bien a las características de dos de las tres variables implicadas en la situación —puntuación en el *ítem* y situación de medida en el diseño—. Esta característica conlleva que la tercera variable involucrada —puntuación en la prueba— se deba categorizar de acuerdo con algún criterio, teórico o estadístico, lo que supone una problemática paralela a la existente con otros procedimientos y que comentaremos más adelante.

Otra característica de interés de estos modelos para adecuarse a la situación que nos ocupa es la posibilidad de establecer ceros estructurales —teóricos— en las celdillas que se desee (Agresti, 1984). Esto es especialmente importante si se tiene en cuenta que la situación del pretest, en una buena parte de las posibilidades de aplicación, será razonable, suponer que, salvo efectos de adivinación, los sujetos carecerán de conocimientos en el área de medida que nos ocupa y, por tanto, se le puede fijar un valor teórico de cero en las celdillas correspondientes.

TABLA 1

RESUMEN DE METODOS PRODUCTIVOS PARA EL ANALISIS DE LA SENSITIVIDAD INSTRUCCIONAL

Indices basados en proporciones	— Diferencias Pretest/Postest	— Cox y Vargas (1966): Brennan y Stolurow (1971); Roudabush (1973); Kosecoff y Klein (1974).
	— Basado en ANOVA.	— Herbig (1975-1976).
	— Diferencias entre grupos independientes	— Levin y Marton (1971); Marton (1973); Klein y Kosecoff (1976); Brennan (1972).
	— Corr. parcial *ítem*-criterio	— Darlingthon y Bishop (1977).
Aproximaciones	— Corr. *ítem*-cambio	— Saupe (1966).
Correlacionales	— Corr. *ítem*-toal de grupos combinados.	— Helmstadter (1972); Haladyna (1974).
Regresion múltiple paso a paso		— Millman (1974).
Aproximación bayesiana		— Helmstadter (1974).
Metodos basados en rasgo latente		— Van der Linden (1981)

Sobre esta base hemos propuesto la aplicación que se recoge en la tabla 2. Como notación empleamos: I *(ítem)*, R (Pretest-Postest) y P (Nivel de habilidad en la prueba). Los tres efectos de asociación de pares de factores encuentran un claro referente teórico. La asociación entre la puntuación en la prueba y la situación (Pre-post) nos informa respecto a la sensibilidad de la prueba en su conjunto para la detección del cambio instruccional. Es decir, la relación coherente esperable es una menor puntuación en la prueba en el pretest que en el postest y siempre de forma significativa. Aquí cabe también señalar que los criterios teóricos de división o categorización de la puntuación en la prueba son mucho más eficaces y sensibles que los estadísticos. Estos últimos al tenerse que realizar sobre la distribución imponen una restricción en el nivel de asociación posible.

TABLA 2

DISPOSICION TRIDIMENSIONAL PARA ANALISIS DE ELEMENTOS CON EL MODELO LOG-LINEAL

N_3		A		B	
		Grupo 1		Grupo 2	
N_2		1	0	1	0
N_1	1				
	.				
	i				
	.				
	n				

N_1 = Nivel de Habilidad

N_2 = Ácierto-Error en el *ítem*.

N_3 = Nivel por Diseño Apto/No-apto Pretest-Postest... (A.B.)

Por su parte, la asociación entre la puntuación en el *ítem* y en la totalidad de la prueba aporta información sobre la discriminación del elemento. Pero es preciso tener en cuenta que la discriminación así evaluada nos informa sobre un «promedio» de las discriminaciones en el pretest y en el postest, con lo que puede confundir los efectos que estén presentes. Así, para valorar la discriminación por sí misma se debería tomar las dos tablas separadas para cada una de las dos situaciones de diseño.

Finalmente, la asociación de la puntuación en el *ítem* y la situación nos ofrece información sobre la sensitividad instruccional del elemento. En este sentido, resulta deseable, genéricamente hablando, que el modelo saturado que incluye las tres asociaciones —IR, IP, PR— alcance un nivel de significación satisfactorio.

Por lo que se refiere a la interacción, o asociación combinada, de los tres factores, es de muy difícil interpretación aun en el caso de que las asociaciones entre los tres pares de variables resultaran significativas. Así, el hecho de que una interacción significativa indique que las asociaciones entre las variables se modifican de acuerdo con los niveles de la tercera de las mismas (Agresti, 1984; Dillon y Goldstein, 1984), únicamente es interpretable con la inspección cuidadosa de la tabla de frecuencia de triple entrada.

Las limitaciones más sustanciales que presenta este procedimiento se pueden resumir en las siguientes:

— La determinación del número de categorías en la puntuación de la prueba plantea una problemática equiparable con los procedimientos. *Chi*-cuadrado para evaluar el sesgo de los *ítems* (Scheuneman, 1975, 1979; Nungesterm, 1977; Camilli, 1979) por lo que se pueden aplicar aquí las recomendaciones que se han recogido en otra parte a este respecto (Jornet, 1987).
— La utilización para categorizar la puntuación en la prueba de criterios estadísticos o teóricos conlleva un conjunto de opciones diferentes que pueden afectar seriamente a los resultados. Sin haber realizado aún un estudio empírico suficiente, pensamos que la utilización del punto de corte como mecanismo de

categorización es la alternativa más congruente, pues nos ofrece información de la sensitividad instruccional de los diversos *ítems* referida al punto crítico respecto del que se van a tomar las decisiones finales.

— Aunque en menos medida que otros procedimientos, como los de rasgo latente, este acercamiento que proponemos requiere un nivel de muestra amplio, en torno a los 200 sujetos. En cualquier caso, es preciso realizar un estudio sobre las implicaciones que el tamaño muestral tiene sobre el procedimiento.

Por último, cabe señalar que la proposición de este procedimiento no es excluyente respecto a las propuestas basadas en Rasgo Latente (Van der Linden, 1981) y Logit (Van der Flier, *et al.*, 1984). En concreto, parece muy adecuado el establecimiento de un procedimiento iterativo con cualquier método como propone Van Der Flier, *et al.* (1984). Así, aunque esta estrategia requiere disponer de programas de ordenador y tiempo de proceso suficientes y, en el caso de seguir la recomendación de categorización apuntada previamente —respecto al punto de corte—, se complicarían extraordinariamente las herramientas que se deben construir para un manejo fluido por el usuario, consideramos que la mejora en la iteración rentabiliza cualquier esfuerzo que se deba hacer al respecto. Finalmente, en la situación actual entendemos que se debe realizar mayor investigación sobre la aplicabilidad comparativa y diferencial de cada uno de ellos para los diversos tipos de pruebas.

REFERENCIAS BIBLIOGRAFICAS

AGRESTI, A. (1984): *Analysis of ordinal Categorical Data.* John Wiley, Nueva York.

BERK, R. A. (ed.) (1984): *A guide to criterion-referenced test construction.* The Johns Hopkins University Press, Baltimore.

BRENNAN, R. L. (1972): «A generalized upper-lower *ítem* discrimination index». *Educational and Psychological Measurement, 32,* 289-303.

BRENNAN, R. L. y STOLUROW, L. M. (1971): *An empirical decision process for formative evaluation.* Paper presented at the Annual meeting of the American Education Research Association, febrero, Nueva York (Editado como: Research Memorandum nº 4. Harvard CAI Laboratory. Cambridge, Mass).

CAMILLI, G. (1979): *A critique of the chi-square method for assessing item bias. Unpublished paper, Laboratory of Educational Research,* University of Colorado, Boulder.

COX, R. C. y VARGAS, J. S. (1966): *A comparison of item Selection techniques for norm-referenced and criterion-referenced tests.* Paper presented at the Annual meeting of the National Council on Measurement in Education, febrero, Chicago.

DARLINGTON, N. B. y BISHOP, C. H. (1966): «Increasing test validity by considering inter-item correlations», *Journal of Applied Psychology, 50,* 322-330.

DILLON, W. R. y GOLDSTEIN, H. (1984): *Multivariate Analysis. Methods and applications.* John Wiley, Nueva York.

HALADYNA, T. M. (1974): «Effects of different samples on item and test characteristics of criterion-referenced tests». *Journal of Educational Measurement, 11,* 93-100.

HELMSTADTER, G. C. (1972): *Comparison of traditional item analysis selection procedures with those recommended for tests designed to measure achievement following performance = + −oriented instruction.* Paper presented at the Annual meeting of the American Psychological Association, septiembre, Honolulu.

HELMSTADTER, G. C. (1974): *A comparison of bayesian and traditional indexes of test item effectiveness.* Paper presented at the Meeting of the National Council on Measurement in Education, abril, Chicago.

HERBIG, M. (1975): «Zur vortest-nachtest-validierung lehrzielorientierter tests». *Zeitschrift fur Erziehungswissenschaftleiche Forschung, 9,* 112-126.

— (1976): «Item analysis by use in pretest and post-tests: A comparison of different coeficientes». *Programmed Learning and Educational Technology, 13,* 49-54.

JORNET, J. M. (1987): *Una aproximación teórico-empírica a los métodos de medición de referencia criterial.* Tesis Doctoral no publicada, Universidad de Valencia.

JORNET, J. M. y SUAREZ, J. M. (1987): *La sensitividad instruccional: una característica métrica para los elementos de los tests de Referencia Criterial (TRC).* En prensa.

KLEIN, S. P. y KOSECOFF, J. B. (1976): «Issues and procedures in the development of criterion-referenced tests», en MEHRENS W. A.: *Readings in Measurement and Evaluation in Education and Psychology.* Holt Nueva York, 276-293.

KOSECOFF, J. B. y KLEIN, S. P. (1974): *Instructional sensitivity statistics appropiate for objectives-based test items.* Paper presented at the Meeting of the National Council on Measurement in Education, abril Chicago.

LEVIN, L. y MARTON, F. (1971): *Prouteory och Proukonstruktion.* Almquist and Wiksell, Estocolmo.

MARTON, F. (1973): «Evaluaringdteory och metodik», en HANDEL, G.; HOLMSTROM, L. G. y THONSON, O. B.: *Universitetsundervisning.* Student-Litterature, Malmoe.

MILLMAN, J. (1974): «Criterion-referenced measurement», en POPHAN, W. J.: *Evaluation: Current Applications.* McCutchan, Berkeley, Ca.

NUNGESTER, R. J. (1977): *An empirical examination of three models of item bias.* (Doctoral Dissertation. Florida State University). Dissertation Abstracts International, 38, 2726A. (University Microfilms Nº 77-24, 269).

ROID, G. H. y HALADYNA, T. M. (1982): *A technology for test-items writing.* Academic Press, Nueva York.

ROUDABUSH, G. E. (1973): *Item selection for criterion-referenced tests.* Paper presented at the Annual meeting of the American Educational Research Association, febrero, Nueva Orleans.

SAUPE, J. L. (1966): «Selecting items to measure change». *Journal of Educational Measurement, 3,* 223-228.

SCHEUNEMAN, J. D. (1975): *A new method for assessing bias in test items.* Paper presented at the Annual meeting of the American Educational Research Association, abril, Washington. (ERIC Document Reproduction Service nº ED 106 359).

— (1979): «A new method for assessing bias in test items». *Journal of Educational Measurement, 16,* 143-152.

VAN DER FLIER, H.; MELLENBERGH, G. J.; ADER, H. J. y WIJN, M. (1984): «An iterative item bias detection method». *Journal of Educational Measurement, 21,* 131-145.

VAN DER LINDEN, W. J. (1981): «A Latent Trait Look at Pretest-Postest Validation of Criterion-referenced Test Items». *Review of Educational Research.* 51, *3,* 379-402.

J. M. JORNET y J. M. SUAREZ. Universidad de Valencia. Facultad de Filosofía y Ciencias de la Educación. Avda. Blasco Ibáñez, 21. 46010 Valencia (España).