Revista Investigación Educativa - Vol. 8 - n.º 16 - 1990 (P. 551-559)

ALGUNAS NOTAS DE REFLEXIÓN METODOLÓGICA ACERCA DEL ESTUDIO DE DISTRACTORES Y EL SESGO DE ÍTEMS EN TESTS EDUCATIVOS Y PSICOLÓGICOS

por
Jornet J. M. y Suárez J. M.
Dpto. M.I.D.E. Universitat de València

0. INTRODUCCIÓN

Una de las limitaciones más frecuentes señaladas para los tests desarrollados sobre modelos de carácter normativo es la baja utilidad de la información que aportan para el diagnóstico al sintetizar la ejecución mostrada por el sujeto en la prueba en una única puntuación, lo cual se basa en la concepción teórica que origina el test. Así, utilizar los tests como unidades diagnósticas capaces de, con su interpretación, guiar la selección de unidades de intervención es un aspecto no sólo no previsto, aunque muy deseable, sino también prácticamente imposible en la mayoría de casos. Ello en definitiva, es una manifestación de los bajos niveles de Validez de Constructo con que se han venido generando las pruebas. El énfasis de los especialistas en medición se pone cada vez más sobre la necesidad de determinar estrategias que permitan desarrollar los tests sobre una sólida base, su Validez de Constructo, así como establecer planes de investigación que aporten evidencias acerca de la misma (BARTOLOMÉ, 1982; MATEO, 1988; LINN, 1988).

Dentro de las corrientes más recientes de construcción de pruebas en Educación y Psicología, el análisis de los errores cometidos en los ítems se ha convertido en una evidencia de Validez, componente adicional en la medida del constructo teórico a que esté destinado el mismo. Son pues estudios que afectan, en definitiva a la Validez de Constructo del instrumento de medida, aunque dependiendo del tipo de aproximación, en ocasiones se reseñan como aportación a la Validez bajo otras acepciones. En este marco, el estudio de los distractores se integra, pues, como una aproximación para la mejora de las características métricas internas del test.

Asimismo, desde una acepción de Validez Criterial, el estudio del comportamiento del test en general y de los ítems, en particular, respecto a su funcionamiento en relación a variables diferenciales o diferenciadoras ha constituido otro foco de interés. Esta aproximación identificada como análisis del sesgo alude a la mejora del test en relación a variables externas al mismo, generalmente definidoras de características de subpoblaciones.

En el ámbito educativo estos enfoques se encuentran parcialmente ligados a los análisis de Dominios Educativos, tanto para la elaboración de tests criteriales, tests asistidos por ordenador o la evaluación de programas, entre otros. Sin embargo, también hay que decirlo, este tipo de aportaciones todavía no se ha extendido. Pasamos, pues, a realizar algunas anotaciones en torno a estas cuestiones.

1. ALGUNAS NOTAS ACERCA DEL CONCEPTO Y APROXIMACIÓN AL ANÁLISIS DE LOS DISTRACTORES

El término distractor —como referente de alternativas de error en un ítem— si bien ha estado presente desde hace mucho tiempo en la construcción de tests psicológicos y educativos ha cobrado nuevo interés en las últimas décadas a partir de trabajos parciales y referencias a sus usos tanto desde autores que desarrollan su labor en el ámbito de la Evaluación Criterial, como desde la Psicología Cognitiva. Este interés, fundamentado en el tema de validación, se ha debido a la concepción de los «distractores» como elementos que deberían ser aprovechados en dos vertientes:

a) meramente psicométrica, como elemento de confusión para el sujeto, de forma que una elección correcta esté avalada por un proceso de decisión entre opciones que se puedan escalar por su plausibilidad y,

b) unos usos diagnósticos, apoyados sobre la validez de constructo, de forma que tanto en la elección correcta como en el error cometido en un ítem de elección múltiple, puedan haber pautas que hagan posible una interpretación procesual o funcional de las respuestas individuales.

Estas dos perspectivas, no es extraño pues, que se estén desarrollando en los ámbitos mencionados anteriormente —tanto para tests que midan variables sujetas a interpretaciones normativas como criteriales— y lo hagan bajo el denominador común de buscar pautas para interpretaciones adicionales a la mera puntuación global en una prueba. Este tipo de estudios, si bien sería más aconsejable realizarlos desde planteamientos experimentales en fases iniciales de construcción de tests, pueden desarrollarse —y creemos que es útil hacerlo— para pruebas ya construidas que han mostrado su calidad psicométrica y utilidad diagnóstica, como un medio de enriquecer su uso, aportando evidencias de su validez.

Sin embargo, pese a la relativa complejidad de las hipótesis implicadas, para el estudio de los distractores no se han planteado procedimientos más allá de su mero análisis frecuencial, lo cual conlleva un trabajo muy pormenorizado por ítem y, sin referentes o criterios genéricos, integrados en modelos de medición, que permitan

valorar adecuadamente las hipótesis que se hayan planteado respecto a cada ítem o en el conjunto del test. Con todo, siendo que el aprendizaje como proceso —y el rendimiento como su expresión de producto— es conceptualizable como multidimensional y no necesariamente lineal (DE LA ORDEN, 1985), el análisis de los distractores puede constituir un apoyo para dotar a los tests de rendimiento de posibilidades interpretativas, dado que en definitiva supone profundizar en el conocimiento del constructo teórico con una intención claramente aplicada. En este sentido, estimamos que en Educación es necesario no sólo ahondar en los tests de que actualmente disponemos, en este tipo de estudios, sino también apoyarlos en otros enfoques complementarios.

2. ALGUNAS NOTAS RESPECTO AL CONCEPTO Y PROCEDIMIENTOS DE ANÁLISIS DEL SESGO DE LOS ÍTEMS

El análisis del sesgo de los ítems supone un anclaje adicional al enfoque anterior. Para considerar el sesgo de los ítems partimos de dos nociones básicas:

a) La noción de que la ejecución en un ítem puede estar sujeta a otras fuentes de variación más que a diferencias en el constructo de interés y,

b) El supuesto de que estas fuentes extrañas de variación influyen en la ejecución de forma sistemáticamente diferente para algunos subgrupos identificables.

¿Qué entendemos por sesgo de los ítems? Se han aportado diversas definiciones (CLEARY y HILTON, 1968; ANGOFF, 1972, 1982; SCHEUNEMAN, 1975; PINE, 1977; SHEPARD, 1982; HAMBLETON, 1984; ORDEÑANA, 1987) y todas ellas coinciden en el análisis de la interacción ítem/grupo para los sujetos que presentan el mismo nivel de habilidad, como el elemento claro de determinación de sesgo. Así, se considera que un ítem está insesgado si: a) Los ítems se afectan por las mismas fuentes de variación en dos subpoblaciones y, b) Entre los sujetos que tienen el mismo nivel de habilidad en el constructo propiamente medido por el test, las distribuciones de fuentes de variación irrelevantes son las mismas para ambas subpoblaciones.

A diferencia del estudio de distractores, para la operacionalización del cálculo del sesgo se han realizado múltiples propuestas que pasamos brevemente a reseñar: a) La metodología Delta-plot (ANGOFF,1972, 1982), basada en el método de Escalación Absoluta de THURSTONE de 1925, supone que las dos distribuciones de los índices de dificultad para dos grupos distintos se ordenarán de forma independiente para cada uno de ellos. Fundamentalmente es una técnica gráfica, aunque se han desarrollado diversos procedimientos estadísticos como los de COFFMAN (1961, 1963), CARDALL y COFFMAN (1964), PLAKE y HOOVER (1979, 1980), así como se han propuesto algunas variaciones metodológicas como las de ECHTERNACHT (1974), SINNOT (1980) o RUDNER, GETSON y KNIGHT (1980). En este mismo contexto, JENSEN (1980) propone un *«Análisis del decremento de Delta»*, basado en la descomposición de la varianza ítem por grupo en dos

componentes: ordinal y disordinal de la interacción. Aunque su definición es muy clarificadora en orden a conceptualizar el sesgo, no parece existir razones que aboguen a favor de su aproximación respecto a las propuestas de ANGOFF. En cualquier caso la facilidad e interpretabilidad del método parecen razones suficientes para su consideración.

b) Otras opciones para la investigación del sesgo de los ítems son las desarrolladas a partir de la *técnica Ji-Cuadrado*. SCHEUNEMAN (1979) propone un cálculo de Ji-Cuadrado teniendo en cuenta únicamente los aciertos al ítem, lo cual ha conllevado diversas criticas (BAKER, 1981). NUNGESTER (1977) y CAMILLI (1979) adoptan un Ji-Cuadrado Total que incorpora al cálculo tanto las respuestas correctas como las incorrectas. Las ventajas de ambos acercamientos residen en el mínimo de requerimientos que imponen por ser una prueba Ji-Cuadrado, así como su relativa facilidad de cálculo e interpretación. Un problema común a ambos acercamientos reside en el establecimiento de un criterio coherente para la determinación de los niveles de habilidad. Por otra parte, de acuerdo con los trabajos de MARASCUILO y SLAUGHTER (1981) las hipótesis probadas en ambos acercamientos no son comparables por verificar aspectos diferentes. Así, en Ji-Cuadrado Total se comprueba si la diferencia de proporciones para los grupos en cada intervalo de habilidad se aparte de cero, mientras que en el Ji-Cuadrado de Aciertos se comprueba si existe alguna diferencia entre los grupos que se aparte de cero, considerando que estas diferencias son iguales para todos los intervalos de habilidad.

c) En los *modelos IRT* se han propuesto los siguientes procedimientos (ORDEÑANA, 1987): Comparación de las ICC, tests de igualdad de parámetros y comparación del ajuste datos/modelo. En este sentido, se han desarrollado diversos estudios comparativos respecto a la funcionalidad de estos métodos y todos ellos «...coinciden en la superioridad del modelo logístico de tres parámetros (IRONSON y SUBKOVIAK, 1979)»(ORDEÑANA, 1987, pág., 186). Sin embargo, comparativamente con otras aportaciones clásicas no parecen estar tan claras sus ventajas. El soporte teórico-matemático es sin duda más sólido, pero sus requerimientos mucho más costosos.

Así, a partir de estudios comparativos (RUDNER y CONVEY, 1978; RUDNER, GETSON y KNIGHT, 1980), con datos simulados, mixtos y/o empiricos, las conclusiones iniciales que pueden extraerse es que los métodos Ji-Cuadrado —especialmente el de Camilli— parecen bastante adecuados apara el análisis del sesgo con muestras medias y pequeñas (aproximadamente N> 1000) y un número de niveles de habilidad medio (alrededor de S; a medida que aumentan los niveles de habilidad decrece la precisión de Ji-Cuadrado). Asimismo los métodos de Ji-Cuadrado total y Delta corregido presentan bastante relación con los resultados que ofrecen las opciones IRT (INTASUWAN, 1979; SHEPARD, CAMILLI y AVERILL, 1980).

3. REFLEXIONES EN TORNO A ESTOS CONCEPTOS EN TESTS EDUCATIVOS

Estas aproximaciones que se pueden considerar adicionales en el análisis de elementos estimamos que pueden tener un gran interés en la construcción y/o adaptación de tests educativos en tanto en cuanto suponen evidencias de validez y, por ello, aportan elementos de mejora en el uso de los mismos. En cualquier caso, en su utilización es conveniente integrar algunas notas que pasamos a comentar a continuación.

1. En el análisis de distractores una opción de trabajo que, aunque costosa, parece aportar resultados de interés es la relativa a su estudio frecuencial tanto en el global de la muestra de normalización, como sobre subpoblaciones establecidas para niveles de habilidad en la prueba y su contraste mediante Ji-Cuadrado o mediante cualquier otra prueba que facilite el contraste de hipótesis. Esta estrategia supone una profundización en el análisis del parámetro de discriminación en ítems de elección múltiple, aspecto éste no suficientemente atendido en los modelos de medición.

Las hipótesis implicadas, para cada uno de los ítems, desde un punto de vista métrico, se pueden sintetizar en el planteamiento que realizamos a continuación. Las alternativas de respuesta en un ítem deberían escalarse de acuerdo a su plausibilidad, configurando un patrón que debería replicarse tanto para el grupo normativo, como para grupos de rendimiento establecidos en función de la puntuación total de la subprueba. Todo ello, claro está, siendo función de la dificultad y capacidad discriminativa del ftem. Un desajuste en este punto, bien tiene una explicación meramente psicométrica como error sistemático, bien es expresión de la influencia de variables en la respuesta de los sujetos relacionadas con el proceso de ejecución, o lo que es lo mismo, es un indicador de aspectos funcionales del constructo teórico que pretende medir una prueba.

En este planteamiento se deben tener en cuenta, asimismo, algunas hipótesis relativas al concepto de adivinación o acierto por azar. LORD (1952) analizó la influencia del número de alternativas de respuesta en la adivinación, sus efectos sobre la estimación del indice de dificultad de los ítems y sus consecuencias para la fiabilidad de la prueba. Teniendo en cuenta sus resultados, podríamos señalar que el fenómeno de la adivinación se puede, pues, conceptualizar en torno a dos dimensiones: a) la dificultad del ítem y, b) la incertidumbre en las elecciones de respuesta. Se puede hipotetizar acerca de una función nomonótona de la adivinación respecto a estas dos dimensiones (si no diferenciamos errores de omisiones, lo cuál complejizaria el modelo), de manera que la máxima adivinación se puede localizar precisamente para los ítems modales donde se identifican a el máximo de incertidumbre. Así, este hecho deberá tenerse en cuenta, con independencia de los criterios que se establezcan para la corrección por adivinación (CHOPPIN, 1988) tanto para la selección de elementos (considerando que los efectos de adivinación afectarán más a ítems medios) como en el mismo análisis de distractores, donde ello deberá comprobarse.

2. En el análisis del sesgo de los ítems todas las aproximaciones han partido de considerar únicamente una estructura de acierto/error en el ítem (puntuación 1/0). En ítems de elección múltiple este planteamiento, aunque también tiene interés, sin embargo, no es suficiente. La extensión del análisis del sesgo a los distractores puede ayudar a conocer mejor los componentes de respuesta en relación a variables externas al test y relacionadas con el constructo medido. En esta línea se han realizado algunas propuestas, entre ellas las de VEALE y FOREMAN (1976) y SCHEUNEMAN (1982) —basadas en Ji-Cuadrado— y FRARY y GILES (1980) —desarrollada desde el modelo de RASCH—. En este sentido nuestra experiencia es coincidente con la señalada por SCHEUNEMAN (1982) que indica que este procedimiento puede suponer un apoyo para determinar mejor el comportamiento de los ítems, aunque no es un camino seguro para identificar las causas del mismo, las cuáles, estimamos sólo pueden esclarecerse a partir de aproximaciones experimentales.

3. Sin embargo, aunque no sea la solución al problema planteado en el punto anterior, el enfoque combinado de un análisis distractores/sesgo no debe limitarse a una mera aplicación del estudio del sesgo sobre los distractores, sino que éste debe integrarse en una estrategia más general de profundización en el análisis del parámetro de discriminación del ítem, incluyendo como componentes: a) el análisis de distractores teniendo en cuenta los niveles de habilidad, descrito en el punto (l) de estas consideraciones y, b) el análisis del sesgo para cada una de las alternativas en relación a variables diferenciales, mediante Ji-Cuadrado —u otro procedimiento—. En definitiva, no se trata de realizar un estudio sobre los errores sino de incluir éstos, diferenciados, como otras opciones de actuación que tienen sentido dentro del diseño global de la tarea. Así, estas alternativas podrán estar afectadas por diversas variables interfirientes que pueden provocar sesgo en la respuesta de los sujetos. En esta reunión presentamos algunos trabajos en los que se integran las dos vertientes.

4. Con independencia de las consideraciones técnicas apuntadas respecto al concepto de sesgo de los ítems, como punto final de estas notas, nos gustaría plantear algunas cuestiones respecto a su papel en los Tests Educativos. El concepto de sesgo deviene de la teoría del Rasgo psicológico, de forma que supone que éste debe estar presente en todos los sujetos pertenecientes a la misma población y asume que las diferencias individuales serán únicamente de nivel. En este sentido, el supuesto métrico inmediato a este planteamiento es que el test internamente debe producir en la población una misma ordenación respecto a sus ítems, por lo que las discrepancias entre las distribuciones de dos ítems en dos subpoblaciones con el mismo nivel de habilidad total, se entienden como sesgo (todo ello, supuesta y comprobada la unidimensionalidad del rasgo). De esta forma, para que un test pueda ser utilizado adecuadamente debe ser independizado del sesgo.

Ahora bien ¿estos mismos supuestos se pueden considerar características asimilables a la definición del rendimiento educativo como constructo? Si partimos de que la educación es una actividad intencional orientada a producir cambios en los sujetos de acuerdo con algún principio u objetivo y, todo ello, utilizando diversas

posibilidades metodológicas de acción, tendríamos que concluir que no. Porque, precisamente, el sesgo puede ser expresión de la Educación recibida —por sus objetivos y/o metodologías o por la influencia de variables contextuales o ambientales— por una determinada subpoblación. De forma que no nos parece evidente y generalizable que todos los tests educativos deban ser independizados del sesgo de sus ítems; el sesgo, en los tests educativos, debe entenderse más como una característica ligada al objeto, utilización del test, etc., entendiendo que deberá ser la comunidad investigadora la que, en cada caso en relación directa con el contexto social en que está inmersa, tendrá que definir qué sesgo es indeseable —e independizar al test del mismo— y qué sesgo puede ser deseable —como expresión de información diagnóstica a utilizar en un sistema más específico de intepretación de las puntuaciones individuales en los tests—.

En este sentido permitasenos plantear algún ejemplo que ilustre esta reflexión. En primer lugar, supongamos que estamos construyendo una prueba que pretende valorar el «potencial de aprendizaje»; realizamos un análisis del sesgo de los ítems y encontramos varios de ellos segados respecto a la variable sexo. Obviamente, si consideramos esta variable como una expresión de la capacidad posible de aprendizaje con independencia de las influencias culturales o introducidas por la educación formal, no podemos admitir una prueba como válida con este tipo de sesgo, por lo que deberemos intentar independizarla del mismo. Otro supuesto, puede ser el relativo a una prueba que intente medir el nivel de conocimientos adquiridos en Matemáticas, por ejemplo en Primer Ciclo de E.G.B.; como primer acercamiento, supongamos que el uso que se le pretende dar a la prueba es fundamentalmente de carácter diagnóstico, orientado no sólo a evaluar el nivel adquirido, sino también a identificar posibles áreas de dificultades. En el proceso de construcción/adaptación, identificamos que algún ítem presenta sesgo respecto a un grupo minoritario y que, precisamente ese sesgo es explicable desde carcaterísticas culturales de definición de dicho grupo. Considerando el uso descrito, estimamos que la interpretación puntual de dichos ítems pueden, precisamente, aportar un elemento adicional en el diagnóstico de forma que independizar al test de ese sesgo supondría eliminar/perder una información que nos ayudaría a modular las interpretaciones derivadas de dicha prueba. Ahora bien, supongamos que este test se pretende utilizar no con una orientación diagnóstica, sino como una prueba meramente de nivel en un esquema educativo que requiriera pruebas de «certificación» de final de ciclo aplicables a un conjunto de población; en este supuesto, ¿sería necesaria o conveniente la independización de dicho sesgo?; entendemos que sí, dado que de mantener el test en condiciones de sesgo éste no supondría un patrón de evaluación igualmente justo para todas las subpoblaciones identificables, por lo que no sería válido. Obviamente, las soluciones que hemos apuntado para cada una de las situaciones son discutibles, pero es precisamente es ese carácter relativo del sesgo lo que deseamos resaltar en estas últimas líneas. Desde esta óptica nos parece de gran interés el análisis del sesgo en los ítems de los tests educativos.

4. BIBLIOGRAFÍA

ANGOFF, W. H. (1972): A technique for the invcstigation of culmral differences. *Ponencia presentada en la reunión anual de la A.P.A.*, Honolulú, Septiembre (ERIC: Servicio de reproducción de documentos, n.º ED 069 686).

ANGOFF, W. H. (1982): Use of difficulty and discrimination indices for detecting item bias. En *R A. Berk (Ed.): Handbook of Methods for detecting test bias*. Johns Hopkins University Press, Baltimore, MD. págs., 96-116.

BAKER, F. B. (1981): A criticism of Scheunemans ítem bias technique. *Journal of educational measurement*, 18., págs. 59-62.

BARTOLOMÉ, M. (1982): *Validez de los Instrumentos de medida, evaluación y assessment. Su adecuación e Importancia en las diferentes sltuadones de prueba*. Material policopiado. Universidad de Barcelona. Barcelona.

CAMILLI, G. (1979): *A critique or the chl-square method for assessing Ítem blas*. Informe no publicado, Laboratorio de Investigación Educativa, Universidad de Coloradp, Boulder.

CARDALL y COFFMAN, W. E. (1964): A method for comparlng the performance of different goups on the Ítems In a tests. Prtinceton, N.J., Educational Testing Service.

CLEARY, T.A. y HILTON, T. E. (1968): An investigation into items bias. *Educational and Psychological Measurement*, 8, págs. 61-75.

COFFMAN, W. E. (1961): Sex differences in raponses to items in aptimde tests. En: *Eighteenth yearbook of the national council on Measurement in education*, págs. 117-124.

— (1963): *Evidence of culturd factors in responses of african Items in an american tests of scholastic aptitude* (research and development reports), New York: College Entrance Examination Board.

CHOPPIN, B. H. (1988): Objetive tests. En J.P. KEEVES: *Educational research, methodology, and measurement: an international handbook*. Pergamon Press, New York.

DE LA ORDEN, A. (1985): Hacia una conceptualización del producto educativo. *Revista de Investlgación Educativa*, 3 (6), págs. 271-283.

ECHTERNACHT, G. (1974): A quick method for determining test biu. *Educational and Psychologicd Measurement*, 34, 271-288.

FRARY, R. B. y GILES, M. B. (1980): Multiple Choice test bias due to answering strategy variation. *Ponencia presentada en la reunión anual de la AERA*, Boston, Abril.

HAMBLETON, R. K. (1984): Criterion-referenced measurement. En T. HUSEN y T. N. POSTLETHWAITE (Eds.): *International encyclopedia of education: research and studies*. Oxford, Inglaterra, Pergamon Press.

INTASUWAN, S. (1979): A comparison of three approaches for determlnlng Item blas in cross national testing. *Tesis Doctoral*, Universidad de Pittsburgh. (Dissertation Abstract International, 40, 261 3A —University Microfilms, n.º 79-24,720—).

IRONSON, G. H. y SUBKOVIAK, M. L. (1979): A comparison methods of assessing item bias. *Journal of Educatlon Measurement*, 16, págs. 209-225.

JENSEN, A. (1980): Blas In mentd testing. Free Press, New York.

LINN, R. L. (1988): Medición Educativa: Algunos problemas y tendencias actuales. En 1. DENDALUCE (coord.): *Aspectos Metológicos de la investigación educativa*. Narcea, Madrid, págs. 149-163.

LORD, F. M. (1952): A thcory of tests scores. *Psychometric Monographs*, 7.

MATEO ,J. (1988): Medición Educativa. Estado de la cuestión en el ámbito español. En 1. DENDALUCE (coord.): *Aspectos Metodológicos de la investigación educativa*. Narcea, Madrid, págs. 164-173.

MARASCUILO, L. A. y SLAUGHTER, R. E. (1981): Statistical procedures for identifying possible

sources of item bias based on chi-square statistics. *Journal od educationd measurement,* 18, 229-248.

NUNGESTER, R. J. (19M): An empiricd examination of three models of item bias. Tesis doctoral, Universidad del Estado de Florida. (Dissertation Abstracts Intemational, 38, 2726A -University Microfilms n.º 77-24, 289).

ORDEÑANA, B. (1987): La Teoría de Respuesta al Ítem: una aplicación al análisis de sesgos de ítems. En 1. DENDALUCE (coord.): *Aspectos metodológicos de la investigación educativa.* Narcea, Madrid, págs. 183-188.

PINE, S. M. (1977): Applications of item characteristic curve theory to the problem of test bias. En: D J. WEISS (Ed.): *Applications of computerized adaptative testing* (RR 77-1). Minneapolis: Departament of Psychology Psychometric Methods Program, University of Minnesota. Marzo, págs., 33-43.

PLAKE, B. S. y HOOVER, H. D. (1980): An analytical method of identifiying biased test items. *Journal of experimental education,* 48, págs., 153-154.

RUDNER, L. M. y CONVEY, J. J. (1978): An evaluation of select approaches for biased item identification. *Ponencia presentada en la reunión anual de la AERA,* Toronto, Marzo. (ERIC, Servicio de reproducción de documentos, n.º ED 069 686).

RUDNER, L. M., GETSON, P. R. y KNIGHT, D. L. (1980): A Monte Carlo comparison of seven biased item detection techniques. *Journal of educational measurement,* 17, págs., 1 - 10.

SCHEUNEMAN, J. D. (1975): A new method of assessing bias in test items. *Ponencia presentada en la reunión anual de la AERA,* Washington, abril. (ERIC, Servicio de reproducción de documentos, n.º ED 106 359).

— (1979): A new method of assessing bias in test items. *Journal of educational measurement,* 16, págs., 143-152.

— (1982): A posteriori analyses of biased items. En BERK, R. A.: Handbook of me*thods for detecting test bias,* The Johns Hopkins University Press, Baltimore, págs., 180-199.

SHEPARD, L. (1982): Definitions of bias. En BERK, R A.: *Handbook of methods for detecting test bias,* The Johns Hopkins University Press, Baltimore, págs., 9-30.

SHEPARD, L. A., CAMILLI, G. y AVERILL, M. (1980): Comparison of six procedures for detecting test item bias using both intemal and extemal ability criteria. *Ponencia presentada en la reunión anual de la NCME,* Boston, abril.

SINNOT, L. T. (1980): *Differences in item performance across groups* (ETS Research Repon, 80-19), Preinceton, N.J.: Educational Testing Service.

VEALE, J. R. y FOREMAN, D. I. (1976): Cualtural variation in criterion-referenced tests: a «global» item analysis. *Ponencia presentada en la reunión anual de la AERA,* San Francisco, abril.