# *EL PROBLEMA DEL SESGO EN LOS TESTS REVISIÓN HISTÓRICA Y CUESTIONES CRITICAS*

Carmen Delgado Álvarez

*RESUMEN*

*Este artículo plantea la importancia del problema del sesgo en los tests y la necesidad de adoptar medidas de protección contra el uso de técnicas psicométricas, que pudieran afectar negativamente a determinados grupos o colectivos sociales. Se señala la importancia que en su evolución ha tenido la crítica generada por colectivos sociales afectados, incidiendo en el impulso que supuso la crítica feminista para la investigación sobre el sesgo de género. Se critica la conceptualización de sesgo y su diferenciación respecto al término "funcionamiento diferencial". Por último se señalan algunas de las principales fuentes de sesgo en los tests.*

*ABSTRACT*

*This brief outlines the problem of bias in tests and the need to find protection against the use of psychometric techniques as they could have a negative influence on specific social groups. The brief also explains how criticism generated by negatively affected social groups has influenced the evolution of this problem and emphasizes the thrust generated by the feminist movement which provoqued the investigation about gender bias. Some critical issues related to present concept of bias and its differentiation regarding the term Differential Functioning are proposed. Some of the main sources of bias in tests are stated.*

## 1. EL SESGO EN LOS TESTS: UNA CUESTIÓN DE "CIERTA" IMPORTANCIA

Los tests psicológicos y educativos se han integrado de tal modo en la cultura occidental que cualquiera de nosotros tiene una alta probabilidad de someterse, o haber sido sometido, a alguno de los que existen en el mercado psicológico. Los profesionales de las Ciencias de la Conducta disponen de una amplia gama de tests, creados en el propio medio o traducidos desde otras culturas que aplican a un gran porcentaje de la población desde la edad preescolar hasta la senectud.

Con las puntuaciones obtenidas en ellos se toman decisiones sobre personas que compiten por un puesto de trabajo o que pretenden incorporarse en algún programa escolar, se hacen peritaciones en tribunales de justicia, se determinan diagnósticos clínicos, y, se hacen asesorías sobre los intereses profesionales de un gran número de estudiantes. Así pues, los tests psicológicos y educativos forman parte de muchos procesos

que implican toma de decisiones, y en general han sido aceptados como un criterio respetable pero además de tomar decisiones que afectan a individuos, también se establecen diferencias entre grupos étnicos, clases sociales, grupos de edad, contexto urbano y rural, hombres y mujeres a partir de los resultados obtenidos con ellos. De este modo, se elaboraron teorías y se constituyeron áreas de conocimiento tradicionales en la psicología, como es el caso de la psicología diferencial.

Sin embargo, los años 70 convulsionaron el mundo de los tests con el fenómeno de contestación social protagonizado por movimientos de derechos civiles y grupos de mujeres, fundamentalmente en Estados Unidos. La denuncia de falta de equidad en los tests por parte de grupos socio-culturales minoritarios alcanzó las esferas de los pronunciamientos legales, y los ámbitos psicométricos resultaron finalmente impactados a través del *Educational Testing Movement* (AMEG, 1973; American Psychological Association, 1974). La incorporación a partir de los años 80 del estudio del sesgo en los tests por parte del *Educational Testing Service* (American Educational Research Association, 1985) convirtió este problema en un capítulo fundamental de la investigación psicométrica. Estudios destinados a detectar posible "trato desventajoso" de los tests confirmaron una situación de desventaja real para diversos grupos minoritarios. Recientemente, el trabajo de Montero (1993) muestra interesantes resultados sobre el efecto de la dominancia lingüística en tests de aptitudes, por parte de estudiantes hispanos en Estados Unidos.

El problema del sesgo, plantea la desventaja con que ciertos sujetos se enfrentan a los tests por el hecho de pertenecer a grupos social o culturalmente en desventaja. Se considera que un test está sesgado contra un grupo, cuando sus miembros tienen menor probabilidad de tener éxito en él, en comparación con sujetos de otro grupo que tienen igual habilidad en el constructo medido. El principal resultado del sesgo en tests psicológicos y educativos es que los miembros de grupos minoritarios, que generalmente cargan ya con unos antecedentes socioculturales desfavorables, son objeto de decisiones que añaden un componente más de desventaja a su historia (Reynolds y Brown, 1984; Ukeje, 1990).

*La sensibilización producida en estos* últimos años hacia la igualdad de los sexos, convirtió el problema del sesgo de género en los tests, en uno más de los muchos que demandan actuaciones dirigidas a contribuir a la aspiración de la igualdad. Sin embargo, mientras en Estados Unidos artículos como los de Lenore Harmon reclamaban ya en los años 70 una revisión de los tests en orden a detectar un posible "trato de favor" para los hombres, creadores de la gran mayoría de ellos, fuera de su país ni se había planteado un estudio sistemático de esa naturaleza.

## 2. DE LA CRÍTICA SOCIAL AL PLANTEAMIENTO DEL PROBLEMA

Si bien es cierto que la investigación sobre sesgo se puede considerar reciente en la historia de la medida en psicología, los antecedentes que revelan la preocupación por este problema se pueden encontrar en el origen mismo de la creación de los tests: el problema del sesgo aparece íntimamente ligado a las técnicas psicométricas desde su creación. Cuando Alfred Binet crea en 1910 su test de inteligencia, encuentra un hecho que llama su atención: los niños de estrato socioeconómico bajo obtienen sistemáticamente peores resultados que los niños de estrato económico alto. Una lectura plana de los resultados podría haber llevado a la interpretación de que los segundos eran más inteligentes que los primeros; una lectura más realista de este mismo resultado era, sin duda, sospechar que los ítemes del test de inteligencia podían estar afectados por alguna variable contaminante que determinaba estas diferencias, ya que no había explicación razonable que permitiera sostener mayor capacidad mental en el estrato socioeconómico alto. Binet consideró que la variable que podía estar afectando los resultados de su test podría ser el entrenamiento cultural, tan diferente en uno y en otro grupo, y por esta razón eliminó ciertas categorías de ítemes en el proceso de validación del instrumento (Binet y Simon, 1916).

En 1912, William Stern observa al igual que Binet, que las diferencias de rendimiento

en los tests entre clases sociales en Alemania son también significativas. Resultados similares se repiten, no sólo en Francia y Alemania, sino también en Bélgica y en Estados Unidos. Este dato que, como se señalaba anteriormente, podría ser interpretado en términos de diferentes capacidades entre las clases sociales, se convierte en un índice de que los tests que favorecen claramente a una clase sobre otra (Stern, 1914). Sin embargo, investigadores posteriores a Binet y Stern responden a la relación entre cociente intelectual (CI) y clase social, planteando un debate sobre el origen genético o ambiental de tales diferencias; pero sin llegar a cuestionar la existencia real de las mismas.

Cattell propone en 1940 el término *"libre de cultura"* para plantear la construcción de tests, en su caso de inteligencia, sin carga cultural en el contenido. Esta propuesta generó un amplio debate entre psicólogos, sociólogos y antropólogos. La concepción de un *"test libre de cultura"* implicaba una contradicción interna, pues parecía imposible concebir un instrumento creado en un medio cultural determinado que no estuviera impregnado por dicho medio, razón por la que el mismo Cattell acabó sustituyendo el término *"libre de cultura"* por el de *"cultura reducida"*, que parecía más ajustado a la realidad. Pero el planteamiento del problema del sesgo como tal, aunque intuído tempranamente, no aparece explícitamente planteado en la literatura psicométrica hasta 1945 con los estudios de Kenneth Eells y Allison Davis, provenientes de la Psicología y la Sociología respectivamente.

La publicación de Eells, Davis, Havighurst, Herrick y Tyler (1951) es la primera en apuntar la posibilidad de que diferencias observadas en puntuaciones de tests, tal vez no reflejen verdáderas diferencias en habilidad, sino que podrían ser una consecuencia del contenido específico de los ítemes. Esta incidencia en el contenido específico de los ítemes de los tests, es un importante antecedente de los estudios contemporáneos sobre sesgo de los ítemes.

Hay dos factores que contribuyen a que el debate iniciado por Eells y cols. se vaya extendiendo: por un lado, el desarrollo de los movimientos de derechos civiles en Estados Unidos obliga a focalizar la atención sobre el acceso igualitario a la educación. Son cuestionados los métodos de selección escolar que determinan el acceso a partir de tests de inteligencia porque, curiosamente, en estos tests los grupos minoritarios (afroamericanos, hispanos, ... y también mujeres) obtienen sistemáticamente peores resultados que el grupo mayoritario. Es decir, aparece un importante factor social que determina el interés por el problema del sesgo, y que motiva el desarrollo de esta línea de investigación en los círculos psicométricos especializados. Por otro lado, desde el punto de vista técnico, el desarrollo de métodos prácticos de detección de sesgo posibilita los trabajos de investigación en torno al tema.

Jensen (1969) aviva la polémica suscitada con una publicación en la que defiende que las diferencias en CI entre blancos y negros, mostradas en los tests, tienen un componente genético que explica el 80% de la varianza intergrupo. Esta afirmación, en un medio políticamente sensibilizado hacia una cultura de *justicia en los tests*, resultó ser un fuerte impulso a la investigación sobre el sesgo de estos instrumentos de medida. Provocó réplicas y contrarréplicas también en el ámbito social, incluyendo el llamamiento de la *Asociación de Psicólogos Negros* para establecer una moratoria de toda evaluación con tests a personas de raza negra, hasta disponer de instrumentos más adecuados. Otros grupos minoritarios también reaccionaron con demandas ante los tribunales de justicia, por decisiones perjudiciales tomadas sobre miembros de sus grupos. El caso de *Diana versus el Consejo de Educación del Estado de California* en 1970 es un ejemplo de este tipo de situaciones (Larry, 1979). En este caso, nueve niños hispanos habían sido asignados a clases de educación especial por haberles aplicado incorrectamente un test de inteligencia que no tenía en cuenta su dominancia lingüística. Cuando este efecto fue corregido, las puntuaciones en CI de estos niños se vieron incrementadas en una desviación típica.

Las diferencias encontradas entre diversos grupos étnicos, mujeres y hombres, clases sociales distintas, etc., plantean la necesidad ética y científica de determinar con rigor si son diferencias reales, o son producto del instrumento de

medida y por tanto debidas al sesgo de los tests. Esto explica por qué investigaciones sobre métodos estadísticos para la detección de sesgo en los tests, aparecen súbitamente en la literatura psicométrica a partir de los años 70, revelando la magnitud de un debate social fuertemente planteado.

Es preciso subrayar que el ámbito de contestación y consecuentemente de investigación, se centra en algunas áreas (inteligencia y aptitudes) y en algunos criterios de pertenencia a grupos (etnia, género, clase social) porque la gente cree que los tests son instrumentos de una sociedad racista (Williams, 1970), y que dan una "injusta ventaja a los hombres y a los miembros de la clase media" (Shepard, 1982, p.12). La relevancia de este cuestionamiento alcanza cotas de tal magnitud, que en algunos casos legales se recoge el mandato de eliminar el sesgo de las mediciones psicológicas (Bersoff, 1982).

## 3. EL IMPACTO DE LA CRÍTICA FEMINISTA

La crítica feminista es en gran parte responsable del impacto del problema del sesgo en el área de la medición psicológica. El artículo de Lenore Harmon (1973) fue el punto de partida para el planteamiento del sesgo de género en muchos tests psicológicos. Harmon cuestionó en aquel artículo la validez de los tests de intereses vocacionales y ocupacionales en los que se podía sospechar *razonablemente*, presencia de sesgo sexual. La *American Personnel and Guidance Association* (APGA), acabó asumiendo este debate y resolvió crear una comisión para estudiar el problema del sesgo sexual en inventarios de intereses vocacionales que circulaban libremente por el mercado psicológico. Consecuencia de esta iniciativa, fue la reformulación de numerosos inventarios que fueron reeditados con las modificaciones que trataban de corregir el sesgo sexual en ellos, mediante los procedimientos recogidos en la *Guidelines for Equal Treatment of the Sexes*, publicada por McGraw-Hill en 1974 (Lewin y Wild, 1991).

El *Educational Testing Movement* planteó que la inclusión de mujeres en las muestras utilizadas para la validación de tests, debía realizarse desde el inicio de la investigación, y no en etapas posteriores como ocurría con muchos tests. El *Educacional Testing Service* (ETS) introdujo cambios específicos en los procedimientos de construcción de tests, como la creación del *ETS Sensitivity Review Process* con normas específicas para la revisión del lenguaje de los tests. El proceso de revisión de la sensibilidad del test, se propone eliminar el lenguaje sexista, racista y potencialmente ofensivo, inapropiado o negativo para cualquier grupo. El impacto de esta revisión condujo a realizar cambios en tests de gran implantación en Estados Unidos, como el SAT. Cruise y Kimmel (1990) estudiaron la evolución en las revisiones sucesivas de los tests verbales del SAT, y encontraron cambios como la eliminación del término genérico "*el*", un incremento del 1% al 16% de ítemes referidos a mujeres, y una representación más balanceada de mujeres y hombres en las últimas revisiones.

Además de la revisión de la sensibilidad del lenguaje, el *Educational Testing Service* adoptó en 1986 el uso sistemático de procedimientos estadísticos para identificar funcionamiento diferencial en los ítemes, y evitar situaciones de desventaja en grupos de mujeres o en grupos minoritarios. Esta práctica se convierte en rutinaria desde entonces, y como señalan Scheuneman y Bleistein (1989) y Cole y Moss (1989), el interés por los problemas de justicia y de sesgo de los ítemes y de los tests, se desarrolla como resultado de los movimientos de derechos civiles y movimientos de mujeres.

Otro ejemplo del impacto de la crítica feminista es "el intento sistemático (del Educational Testing Service) de incluir mujeres en los comités responsables del desarrollo de tests" (Lewin y Wild, 1991, p.590). El incremento de la representación de mujeres en estos comités, es una demanda de la perspectiva del feminismo científico, asumida por el *Educational Testing Service*. Es en los comités donde se decide el contenido, la redacción y la revisión de los ítemes, y la supervisión del test. Ejemplo de este cambio, es el comité de los tests para el examen de graduados, en el que la representación de mujeres pasó del 6% en 1970 al 29% en 1990, o el comité del programa de nivelación avanzada que pasó del 22% al 46% (Lewin y Wild, 1991). La presencia de mujeres en los comités ha facilitado la re-

visión y el cambio de aspectos de los tests relacionados con el género.

El *Código de Prácticas Justas de Medición en Educación* creado por la Junta de Comités de Prácticas de Medición de la American Psychological Association (1988), se añade a la lista de ejemplos que muestran el impacto de la crítica feminista en el área de la medición en psicología. Sin embargo, aunque el sesgo de género ha sido investigado en tests de inteligencia y en campos de aplicación muy concretos (selección escolar y selección de personal principalmente), su aplicación al ámbito de la personalidad y las actitudes ha sido prácticamente nula, si bien actualmente en muchos procesos de selección se tienen en cuenta características de este tipo en los candidatos.

El sesgo en tests de personalidad no ha recibido la misma atención que en los tests cognitivos. La importancia de distinguir verdaderas diferencias y diferencias debidas al sesgo ha producido mucha investigación en psicología; pero como señalan Thissen, Steinberg y Gerrard (1986) es necesario extender esta investigación al campo de la personalidad y las actitudes. Los estudios interraciales de escalas de personalidad no han sido planteados en el paradigma del sesgo, y las diferencias encontradas han sido interpretadas como verdaderas diferencias en las dimensiones de personalidad estudiadas. Sin embargo, la evidencia parece mostrar que el sesgo cultural tiene una mayor probabilidad de estar presente en medidas de personalidad que en medidas cognitivas (Reynolds y Paget, 1983; Reynolds, Plake y Harding, 1983). El sesgo de género que Thissen, Steinberg y Gerrard (1986) encontraron en un inventario de culpa sexual,

> "es muy probable que sea un problema común a muchos cuestionarios de personalidad y actitudes que se están utilizando, y que crean confusión sobre las diferencias encontradas entre los grupos. Los constructores y usuarios de tales escalas, deben investigar esta posibilidad" (p. 126).

Scheuneman y Gerritz (1990) concluyen su trabajo sobre las fuentes de funcionamiento diferencial de los ítemes para hombres y mujeres, sugiriendo

> "la necesidad de balancear algunos aspectos del contenido y de otros aspectos de los ítemes, en orden a reducir los efectos que estas variaciones formales de las propiedades de los ítemes producen, en los resultados de las mujeres y otros grupos minoritarios" (p. 129).

En neuroticismo, por ejemplo, se confirman consistentemente diferencias en función del género, así como en función de la clase social: al igual que la clase social baja puntúa más alto que la clase alta, las mujeres dan puntuaciones más altas que los hombres en más de treinta países diferentes (Francis, 1993). Estos resultados aparecían ya en los estudios de Saville y Blinkhorn (1976) y se han mantenido tanto en investigaciones realizadas con el *Eysenck Personality Inventory* (Simon y Thomas, 1983), como en estudios realizados con distintas formas del *Eysenck Personality Questionnaire* (Corulla, 1988;1989). La forma general de interpretar estos resultados ha sido la afirmación de que tales diferencias existen, lanzando hipótesis explicativas de su origen; pero sin plantear la posibilidad de que tales resultados sean producto del instrumento de medida, aunque investigadores clásicos del problema del sesgo, como Reynolds y Brown (1984) han señalado que

> "puede existir sesgo no sólo en los tests mentales, sino también en otro tipo de test psicológicos, como tests de personalidad, vocacionales, y psicopatológicos" (p. 14).

De hecho, el estudio de cuestionarios como el *Eysenck Personality Inventory* desde la metodología del sesgo, ha mostrado que muchos de los ítemes de Neuroticismo están afectados por contenidos de rol de género que desfavorecen a las mujeres (Delgado, 1994; 1995; Delgado & Martin, 1997a; 1997b).

## 4. EL PROBLEMA DE LOS IMPLÍCITOS O EL ABANDONO DE UN DEBATE INACABADO

El concepto de sesgo implica siempre una diferenciación entre sujetos en función

de su grupo de pertenencia. El grupo, por tanto, es central en los estudios de sesgo y en principio el número de grupos a comparar es ilimitado, aunque se suelen tomar dos a los que se denomina *Grupo Focal y Grupo de Referencia* (para estudios con más de dos grupos, ver por ejemplo, Kim, Cohen & Park, 1995). El *Grupo Focal* es aquel que se considera perjudicado por el test objeto de estudio, y el *Grupo de Referencia*, aquel con el cual se compara. También se utilizan los términos *minoritario y mayoritario* para referirse a ellos, puesto que el *Grupo Focal* suele ser minoritario en tamaño respecto al *Grupo de Referencia*, y lo es siempre en términos de poder social. Si bien la pertenencia a los grupos puede establecerse a partir de cualquier variable sociodemográfica que se considere relevante para detectar significaciones diferentes de las puntuaciones obtenidas en un test (etnia, clase social, género, hábitat rural-urbano, edad, nacionalidad, religión, ...), la investigación sobre sesgo ha estado centrada principalmente en torno a la etnia y el género.

Si la presión social fue determinante para que el problema del sesgo cobrara relevancia en el ámbito de la medición, el protagonismo de ciertos grupos de presión (en este caso, grupos étnicos y grupos de mujeres) determinó de igual modo, que su problemática fuera la más investigada. Este origen *social* de la detección del problema del sesgo, ha hecho que su conceptualización apareciera ligada a otros conceptos ajenos al aspecto meramente técnico, y que su delimitación no fuera tan evidente desde el principio. De ahí que desde las primeras investigaciones, fuera necesario diferenciarlo de otros conceptos afines, y que algunas veces se plantearan controversias entre los principales investigadores. Mertz (1974) y Green (1975), por ejemplo, establecieron ya en sus estudios que un test puede ser usado justa o injustamente, parcial o imparcialmente; pero la imparcialidad no es atributo de un test, del que sólo se puede decir que está sesgado o no está sesgado. Jensen (1980), Reynolds (1982a) y Osterlind (1983), entre otros, comparten esta distinción entre justicia o imparcialidad en la medición, y ausencia de sesgo en los tests. Esta distinción es problemática para otros autores, como Shepard (1982), quien señala cierta artificialidad en ella y viene a decir que es "*el mismo mensaje aunque con distinto cartel*", y que tratar de diferenciar el sesgo definiéndolo como algo propio del test, frente a la parcialidad en su uso como algo ajeno a él, "*es también una burda distinción que han hecho los psicometristas*" (p. 10).

Una primera delimitación del concepto de sesgo en los tests exige, por tanto, su diferenciación terminológica respecto a otros conceptos ético-sociales, como imparcialidad o justicia, de los que innegablemente es heredero. Pero si bien es cierto que es necesario trazar una línea divisoria entre estos términos, también es cierto que no existe esa línea de separación neta ya que, como señala Muñiz (1992),

> "el problema del sesgo viene acompañado de serias implicaciones sociales en el uso de los tests, pues de darse tal sesgo ciertos grupos sociales, clásicamente blancos-negros, mujeres-hombres, pobres-ricos, etc., cualquiera otra partición es posible, sufrirán la consecuencia" (p. 198).

Es cierto que mientras los conceptos de justicia o imparcialidad tienen una connotación ética, social e incluso política, el término *sesgo* ha sido utilizado en el ámbito científico restringiéndolo al problema de la validez de constructo de los tests; pero conviene no perder de vista que esta distinción, es más estratégica que sustantiva y puede "ocultar" los implícitos ideológicos de lo científico (Delgado, 1997).

Jensen (1980), en una posición totalmente contraria a la que se acaba de exponer, habla de "*conceptos inadecuados de sesgo de los tests*" o "*falacias*" (p. 370). Para referirse a los términos de carácter ético-social con que se vincula el problema del sesgo de los tests. En su opinión las, más importantes, por su extensión y alcance, serían:

- *La falacia de la igualdad*: concepción equivocada de sesgo, que se basa en la suposición gratuita de que las poblaciones humanas son esencialmente idénticas o iguales en cualquier rasgo o habilidad que se mida con un test. Se considera que el test está sesgado cuando se obtienen diferentes distribuciones de las

puntuaciones (media, desviación típica, o cualquier parámetro) para las subpoblaciones comparadas. Para que un test se considere no sesgado, deberían ser minimizadas o eliminadas las diferencias estadísticas entre los grupos. Esta es, a juicio de Jensen, una concepción *ideológica* de sesgo que no puede ser aceptada, puesto que asumir que distintas poblaciones (blancos-negros, rurales-urbanos, hombres-mujeres, etc.) son necesariamente idénticas en inteligencia o cualquier otro rasgo *no es más que una falacia ideológica.*

• *La falacia de la dependencia cultural*: se presume que ciertos grupos de la población han tenido experiencias culturales diferentes. Entonces, se considera que el test está sesgado cuando el juicio de expertos declara que el contenido de algún ítem del test, sitúa en posición desventajosa a uno de los grupos de comparación. Es el caso de ítemes de tests de aptitudes como "*¿quién escribió Hamlet?*". Esta también es para Jensen una concepción ideológica del sesgo, pues la cultura se supone un bien universal que se conoce o desconoce, independientemente del grupo de origen.

• *La falacia de la estandarización*: concepción según la cual, puesto que un test ha sido estandarizado sobre una población dada, está necesariamente sesgado cuando se usa con otra población diferente. Sería el caso de algunos tests de inteligencia como Stanford-Binet, Wechsler, etc., que han sido estandarizados con población blanca y son aplicados a población negra. Tampoco esta concepción de sesgo es correcta para Jensen, ya que se puede solucionar este problema haciendo una transformación de escalas cuando se comparan las poblaciones, "*como cambiar de un termómetro Fahrenheit a uno Celsius*" (Jensen, 1980, p. 372).

Los argumentos de Jensen sugieren algunas observaciones tanto de aspectos formales como de contenido. En primer lugar Jensen contrapone el término "*ideológico*" al de "*científico*" para descalificar o invalidar unas concepciones de sesgo, a su juicio inadecuadas. Sin embargo, es obvio que esta contraposición ciencia-ideología en sí misma es una falacia, como bien apunta Sharratt (1993). Si aceptamos la filosofía de la ciencia insaturada por (Kuhn, 1962), el binomio ciencia-ideología es indisoluble y su negación es también una posición ideológica. Tanto, que Kuhn (1970) define el paradigma científico como

> "la constelación de creencias, valores y técnicas que comparte una comunidad dada" (p.12).

La imposibilidad de objetividad en la ciencia, [cuando es negada]

> "hace que quedemos atrapados en nuestras posiciones paradigmáticas, así como en nuestro lenguaje" (Reynolds y Brown, 1984, p. 3).

Por tanto, la primera crítica a las objeciones de Jensen se dirige a los implícitos que subyacen en sus argumentaciones contra las supuestas *falacias* por él denunciadas. Es importante destacar este aspecto en este autor, porque los argumentos con los que hace su crítica son precisamente del tipo que él llama "*ideológico*". Como señala Hilliard (1984)

> "si bien Jensen redujo la posición de las National Education Association y de la Association of Black Psychologist a fulminación verbal y propaganda, él hizo lo mismo (...). Todos sus comentarios en este aspecto, tienen más carácter de propaganda que los que él ha elegido castigar (...) La <<exhaustiva revisión>> de Jensen, ignora datos relevantes"(p. 141).

En segundo lugar, sobre el contenido de la crítica a cada una de las supuestas *falacias* se han planteado diversas críticas que develan el *carácter ideológico* de los argumentos de Jensen. Respecto a la *falacia de la igualdad*, Jensen está asumiendo que la inteligencia es algo objetivable y que los resultados de un test cuantifican lo que los sujetos tienen de *eso* medible, por lo que las diferencias entre grupos no sólo no es prueba de sesgo, sino que es algo totalmente esperable y lógico. En concreto, para el caso de la comparación interracial existe un alto componente genético que determina la superioridad de los blancos. Por tanto, en opinión de Jensen,

calificar de sesgados los tests que muestran estas diferencias y tratar de eliminarlas manipulando los instrumentos de medida, no es ,más que la *falacia de la igualdad.* Sorprende que Jensen ignore que el concepto de inteligencia no es algo universal, desde que Binet acabó diciendo que *inteligencia es lo que miden los tests.* Las definiciones de inteligencia son eminentemente operativas; cada test define operativamente su concepto de inteligencia, y *eso* es lo que mide o pretende medir. Esto ocurre con cualquier constructo que se mida mediante cuestionario. En personalidad, por ejemplo, Thisen, Steinberg y Gerrard (1986) respecto al constructo <<culpabilidad sexual>> señalan que "la medida de <<culpa>> ha sido reducida a una dimensión definida por los ítemes que componen el instrumento" (p. 122) y de haber variado la composición del inventario, modificando la distribución de ítemes que lo componen, las diferencias encontradas entre hombres y mujeres hubieran sido de muy distinto signo. Algo tan simple como sustituir los ítemes relacionados con la infidelidad, por ítemes relacionados con la homosexualidad, hubieran mostrado mayor culpa sexual en los hombres y no en las mujeres. La verdadera falacia, en la que Jensen incurre, está en tratar de darle estatus de *unicidad* a un constructo que se ha definido desde uno de los grupos. Es improcedente proponer *una* conceptualización de un constructo (inteligencia, neuroticismo, ...) a partir de una población, y pretender convertirla en *la* conceptualización, aplicable a cualquier otra (Mischel, 1968; Hampson, 1988).

En lo que se refiere a la *falacia de la dependencias cultural,* Jensen aplica metodología estadística y diseños experimentales tradicionales aún cuando estos procedimientos no fueron diseñados para el análisis de la presencia o ausencia del sesgo cultural. Los supuestos básicos de estos procedimientos implican la no-existencia de cultura. La falacia está una vez más en el mismo supuesto de *unicidad,* ignorando que lenguaje, valores, experiencias, símbolos y reglas, constituyen un universo cultural distinto para grupos diferenciados, y que esta diversidad de antecedentes culturales determina las experiencias y la forma en que se desarrollan las capacidades de los sujetos; es decir, *cualifican* su desarrollo intelectual y psíquico en general. La consecuencia de ignorar la dependencia cultural, es que esta diferencia *cualitativa* se traduce en una falsa diferencia *cuantitativa.* Por tanto, la evaluación de sesgo en el contenido de los ítemes no es un procedimiento subjetivo como pretende demostrar Jensen, sino que la introducción de las variables de contexto es la metodología adecuada para los estudios interculturales, como han demostrado los principales investigadores de este campo (Van de Vijver y Poortinga, 1985; Poortinga, 1986; Poortinga y Van de Vijver, 1987; Van de Vijver y Poortinga, 1991). No olvidemos que, en último término, el problema del sesgo es una cuestión de "*carga cultural*" hasta el punto de cuestionarse la legitimidad de lo que ha sido una práctica habitual en psicología: la traducción de tests, para ser aplicados en medios lingüísticos diferentes. Se considera que el lenguaje, mediador de procesos más complejos, es un elemento definidor de la cultura hasta el punto de considerar su análisis como parte esencial del estudio de las similaridades y diferencias culturales de poblaciones diferentes (Ellis, 1989; Hofstee, 1990; Yang y Bond, 1990).

Por último, en cuanto a *la falacia de la estandarización* ante la argumentación de Jensen de que normas específicas para cada subpoblación complicarían la interpretación de los resultados además de resultar muy costoso, Hilliard (1984) responde que la complejidad y el coste no son criterios precisamente académicos,

> "el problema académico es si una población real ha sido representada adecuadamente (independientemente de lo caro o inconveniente) o si la variedad de poblaciones reales han sido representadas también adecuadamente" (p. 150).

Por otro lado, la diferencia cultural en las poblaciones, es un problema mucho más complejo que *cambiar de grados Fahrenheit a grados Celsius;* es decir, no se puede reducir a un problema de cambio de escala de medida. Una vez más Jensen desconoce la naturaleza de la diferencia cultural. Si se lleva su argumento a una posición extrema, se podría aplicar un test construído con población

bereber, por ejemplo, a una muestra occidental: bastaría con transformar las escalas. Es fácil suponer que los ítemes que pretendan medir la capacidad intelectual de los europeos con problemas sobre el tratamiento de camellos, por ejemplo, los dejarían realmente malparados por mucho que se transformen las escalas. Y es que esta forma de proceder puede ocultar, y en muchos casos oculta, características del test que perjudican a un grupo frente a otro y que no se resuelven con la mera transformación escalar.

Es incuestionable que los problemas derivados de la *dependencia cultural* y de la *estandarización de puntuaciones* a los que Jensen ha otorgado el rango de *falacias*, tienen gran relevancia en la investigación actual. Así, un problema ligado a estas cuestiones es el de la equivalencia de las puntuaciones obtenidas con tests psicológicos traducidos desde otras lenguas y países culturalmente diferentes, lo que ha llevado a cuestionar ya en los años 80, muchas de las conclusiones obtenidas con este tipo de estudios (Candell y Hulin, 1987; Hulin, 1987). La dominancia de los países sajones en la creación de tests psicológicos que se han exportado al resto de países y culturas, es un hecho incuestionable en la historia de la psicología. Instrumentos psicológicos clásicos, tanto en el ámbito de la inteligencia como de la personalidad, han sido vertidos desde su procedencia sajona a las diferentes lenguas de los países que *consumen* psicología. Sin embargo, como señala Ellis (1989)

> "el proceso de traducción, necesariamente introduce el problema de la no-equivalencia de la medida: ¿Es equivalente el test original al test traducido?" (p. 912).

Por tanto, en relación con el argumento de la *falacia de la estandarización* esgrimido por Jensen, lo que verdaderamente constituye una decisión *ideológica* es pretender convertir en dos problemas diferenciados sin conexión entre sí un problema que es único, aunque abordable desde aspectos diferentes. Investigadores que se han interesado claramente por el aspecto meramente técnico del problema del sesgo no han dudado en mantener unida esta dimensión del problema a sus implicaciones éticas, políticas, sociales y económicas. Como señalan Mazor, Clauser y Hambleton (1992),

> "frecuentemente el grupo que tiene ventaja (en la probabilidad de éxito en los ítemes del test), es el grupo mayoritario, o el grupo con ventaja socioeconómica. Por esta razón el problema del sesgo en la medición, que incluye pero no se limita al funcionamiento diferencial de los ítemes, se ha convertido en un importante problema político y legal" (p. 444).

## 5. EL PROBLEMA DEL SESGO, UN PROBLEMA DE VALIDEZ

El principal problema de la detección de sesgo es distinguir verdaderas diferencias entre grupos y diferencias debidas al instrumento de medida. Dentro del instrumento de medida, también se distingue entre el test en su conjunto y los ítemes que lo constituyen. El hecho de que un test contenga ítemes sesgados no implica necesariamente que el test en su conjunto esté sesgado, ya que podrían equilibrarse entre sí los ítemes que desfavorecen a los diferentes grupos de comparación. Se distingue por tanto, entre sesgo de los tests y sesgo de los ítemes; pero independientemente de que se considere el test en su conjunto, o los ítemes individualizados, el término sesgo hace referencia a la validez del instrumento. En una perspectiva integrada sobre el concepto de validez se considera que las distintas categorías con que tradicionalmente se designaban los distintos *tipos de validez* son en realidad "tipos de evidencia dentro de un concepto unitario de validez" (Cole y Moss, 1989, p. 202). Sin embargo, entre estos *tipos de evidencia* sigue siendo un elemento útil para clasificar las definiciones que se han dado sobre sesgo, y así se pueden distinguir definiciones que enfatizan el aspecto de la validez de constructo, predictiva, o de contenido.

Los primeros planteamientos sobre sesgo en los tests abordaban el problema desde la validez predictiva, y en general consideraban que un test se podía considerar sesgado si generaba rectas de regresión diferentes

para los distintos grupos. A partir de la definición de Cleary (1968) que consideraba un test sesgado contra un subgrupo de la población cuando los errores de predicción para sus miembros eran consistentemente distintos de cero, diversos autores propusieron diferentes definiciones de sesgo. Así, autores como Darlington (1971; 1978), Linn (1973; 1979; 1980; 1982; 1984) y Hunter y Schmidt (1976; 1978) interpretaron la definición de Cleary en el sentido de que un test estaría insesgado sólo si los sujetos con igual habilidad (medida con el test) obtienen la misma puntuación en el criterio que se trata de predecir. Otros autores, profundizaron más en los procedimientos de selección de sujetos a partir de las puntuaciones obtenidas con el test (por ejemplo, Thorndike, 1971; McNemar, 1975).

El planteamiento del sesgo en el contexto de la validez de constructo del test surge posteriormente al considerar que la perspectiva de la validez predictiva es una conceptualización parcial y limitada, y que no existe una solución estadística para este problema (Shepard, 1982). El riesgo de convertir el problema del sesgo en una cuestión estadística, hace que algunos autores reivindiquen la necesidad de responder a esta cuestión usando también el debate moral como método (Ellett, 1980), ya que cualquier decisión estadística "es en sí misma una posición de valor" (Shepard, 1982, p. 17).

Desde la validez del constructo se plantea el problema del sesgo en relación con la presencia en el test de *otras variables* que contaminan o alteran las puntuaciones obtenidas en el mismo. Así por ejemplo, Reynolds (1982a) considera que

> "existe sesgo respecto a la validez del constructo cuando un test muestra medir diferentes rasgos hipotéticos (constructos psicológicos) para uno y otro grupo, o cuando mide el mismo rasgo pero con diferente grado de precisión" (p. 194).

Las implicaciones derivadas de esta perspectiva es que de confirmar la existencia de sesgo, por ejemplo en función del género o la etnia de los sujetos, "muchas de las investigaciones de la psicología diferencial deben ser desechadas" (Reynolds, 1982b, p. 200). Este problema coloca en una situación muy difícil las bases de gran parte de la psicología diferencial, en el caso de confirmarse la presencia de sesgo en las escalas utilizadas para sus hallazgos (Reynolds y Brown, 1984).

Desde la validez de contenido se evalúa la pertinencia y representatividad de los ítemes, que constituyen el test, para medir el constructo. En este contexto, se relaciona por tanto el sesgo del test con el contenido específico de los ítemes que lo componen (por ejemplo, Linn y Harnisch, 1981). Estudios como el ya mencionado de Thisen, Steinberg y Gerrard (1986) sobre un cuestionario de culpa sexual, se sitúan en esta perspectiva y muestran sesgo desfavorable para las mujeres, por el imbalance de ítemes referidos a los diferentes tópicos incluídos en el cuestionario.

En suma, si bien el concepto de validez es único, su relación con el sesgo se ha abordado en los diferentes momentos, enfatizando más alguno de sus aspectos. Actualmente el problema del sesgo se plantea como un problema de la validez de constructo del instrumento (Camilli y Shepard, 1994).

## 6. DEL SESGO AL FUNCIONAMIENTO DIFERENCIAL

El concepto de sesgo fue adquiriendo matizaciones a partir de las investigaciones sobre las causas que lo producían hasta el punto de plantear la necesidad de términos diferenciados para referirse a dos aspectos que hasta entonces se incluían en el concepto genérico de sesgo. Si tomamos como referencia los ítemes del test, se consideraba que éstos estaban sesgados cuando tenían distinta dificultad para los grupos de comparación. Como señalan Camilli y Shepard (1994)

> "la idea era igualar a los sujetos en la puntuación total del test, y ver si sujetos comparables de los diferentes grupos también tenían puntuaciones iguales en los ítemes de los test. Si esto no era así, el ítem se consideraba potencialmente sesgado" (p. 15).

Pero este enfoque fue cuestionado por considerar que las medidas de dificultad relativa no constituyen por sí mismas una prueba evidente de sesgo. Se planteó entonces que sólo cuando la fuente de dificultad relativa del ítem es irrelevante para el constructo que pretende medir el test, se puede hablar de sesgo en el ítem. En los demás casos, se trataría de un funcionamiento diferencial del ítem que no necesariamente implicaría sesgo, ya que podría deberse a factores pertinentes para la medida del constructo. Esta es la razón por la que Angoff (1982) sugiere que se reserve una denominación específica para los índices estadísticos que detectan el funcionamiento diferencial, y distinguirlos así del concepto de sesgo del ítem. El término que Angoff propone es *"métodos de discrepancia del ítem"*. Holland y Thayer (1988) proponen un término que ha tenido más éxito que el anterior y que actualmente está ampliamente aceptado: *"funcionamiento diferencial de los ítemes"*, conocido también por las iniciales de su formulación en inglés: *DIF*.

Se puede distinguir entre funcionamiento diferencial del test en su conjunto (DTF) y funcionamiento diferencial de los ítemes (DIF), al igual que se distingue entre sesgo del test y sesgo de los ítemes. Para referirnos al concepto implicado en ambos casos, en contraposición a lo que se ha definido como sesgo, hablamos de funcionamiento diferencial en general y lo dicho respecto al DIF es aplicable al concepto de DTF.

Idealmente, los estadísticos DIF deberían ser utilizados para identificar todos los ítemes que funcionan de forma diferente para grupos diferentes. Pero sólo mediante análisis lógicos posteriores, como la explicación de *por qué* los ítemes parecen ser relativamente más difíciles para un grupo que para otro, que podrá determinar si están o no sesgados. Como señalan Camilli y Shepard (1994), "sólo si la fuente de dificultad diferencial es irrelevante para el constructo medido (...) se puede decir que el ítem está sesgado contra un grupo particular" (p.134).

Esta diferenciación entre sesgo y DIF, teóricamente aceptable y totalmente aceptada por los principales investigadores del tema (Angoff, 1993; Camilli, 1993; Camilli y Shepard, 1994; Donoghue y Allen, 1993) plantea sin embargo un problema en muchos casos difícil de resolver: ¿Cómo identificar la fuente de dificultad diferencial y cómo determinar su relevancia o irrelevancia para el constructo medido? (Breland, 1991; Bridgeman y Lewis, 1991; Mazzeo, Schmitt y Bleistein, 1991). Incluso, determinada la relevancia para el constructo, se puede plantear el problema de la legitimidad o ilegitimidad de su presencia en la escala, como ocurriera con el test de lectura investigado por Scheuneman (1979). La diferencia entre sesgo y DIF, por tanto, no parece tan obvia e indiscutible, como podría pensarse ante la ausencia de debate sobre esta cuestión. Aunque teóricamente la diferenciación no ofrece problemas, en la práctica no parecen tan objetivables las inferencias sobre la ausencia o presencia de sesgo, a partir de los índices de DIF. La obtención de índices de DIF significativos en los ítemes, llevaría a tres situaciones posibles:

• *Fuente de DIF irrelevante para el constructo*: Es el caso en que examinamos los ítemes, la explicación plausible de la dificultad diferencial entre los grupos se atribuye a una variable totalmente ajena al constructo que se pretende medir. En este caso, se puede inferir que el ítem está sesgado.

Por ejemplo, Dorans y Kulick (1983) encontraron que un ítem de la escala verbal del SAT, mostraba mayor dificultad para las mujeres que para los hombres ($\Delta p = .15$). Era un ítem de analogías en el que los sujetos debían identificar pares de palabras que tuvieran entre ellas la misma relación que el par de palabras propuesto. Una traducción del ítem, a efectos meramente ilustrativos, podría ser la siguiente:

TRAMPA-ESQUIVAR: (A) red-mariposa (B) telaraña-araña (C)cebo-pescado (D) lazo-cuerda (E)desvío-acortar

La conclusión de Dorans y Kulick, tras el análisis de las respuestas de los dos grupos, fue que algunas opciones del ítem requerían familiaridad con la jerga propia de la caza y la pesca. Puesto que este conocimiento es irrelevante para el razonamiento verbal, constructo que pretendía medir el test, se podía considerar que el ítem estaba sesgado y que, en general, los ítemes que utilizan términos deportivos resultaban más difíciles para las mujeres que para los hombres. Por esta razón, se considera que el uso del lenguaje deportivo en algunos tests, es fuente de sesgo de género (Camilli y Shepard, 1994).

Este descubrimiento de la interacción entre el lenguaje utilizado y el grupo de pertenencia hizo que se incluyera el *análisis de sensibilidad* de los grupos, como un elemento más a tener en cuenta en los procedimientos de construcción de tests, y así ha sido adoptado por el *Educational Testing Service* (1987). *El análisis de sensibilidad*, que actualmente está dando lugar a la revisión de muchos tests, pretende depurar el lenguaje de las técnicas psicométricas, no sólo para resolver problemas como el del ejemplo anterior, sino también

> "para detectar (...) lenguaje ofensivo y asegurar una representación justa de los roles laborales y los estilos de vida de poblaciones de distinto sexo, raza o etnia" (Berk, 1982, p. 4).

• *Fuentes de DIF relevantes para el constructo*: Es el caso en que, examinados los ítemes, la variable que parece explicar las diferencias en dificultad entre los grupos está relacionada con el constructo que se pretende medir. No se puede inferir sesgo, puesto que las diferencias son atribuibles a una variable relacionada con el constructo medido. En algunos casos, se puede inferir la presencia de un factor secundario, contenido en este ítem específico. Esto es más fácil de detectar cuando aparece en un grupo de ítemes con contenido similar y se reconoce un patrón en ellos. En este caso, los índices de DIF estarían detectando una multidimensionalidad en el test (Dorans y Holland, 1993; Shealy y Stout, 1993b).

Sin embargo, la explicación de la multidimensionalidad propuesta desde la perspectiva DIF no resuelve el problema de la presencia de estos ítemes en el test. Se requiere determinar si esta multidimensionalidad es "*legítima*" o "*ilegítima*" (Camilli y Shepard, 1994) para los propósitos del test. Por ejemplo, en los primeros estudios sobre sesgo Scheuneman (1979) obtuvo un desproporcionado número de ítemes del subtest del lenguaje del *Metropolitan Readiness Test* con índices de funcionamiento diferencial desfavorables para los niños afro-americanos. No ocurría así con los ítemes de los subtests auditivos y visuales. El análisis lógico reveló un patrón de estructuras gramaticales negativas en los ítemes que mostraban DIF, ante las cuales los niños afro-americanos tenían una dificultad relativa mayor. Si bien esta fuente de funcionamiento diferencial era relevante para el constructo, Scheuneman consideró que la multidimensionalidad en este caso era *ilegítima* y recomendó cambiar esos ítemes del test. El argumento era que el uso de las formas negativas, podía considerarse una habilidad del lenguaje que no era esencial para los niños de aquel grado escolar. Investigaciones recientes apoyan la recomendación de esta autora, de tener en cuenta el contexto en que el test es utilizado, para tomar decisiones sobre la pertenencia de este tipo de ítemes (Ellwein, Walsh, Eads, y Miller, 1991; Shepard y Graue, 1993); es decir, sobre la legitimidad de la multidimensionalidad. Esta, por tanto, si bien no puede considerarse equivalente al sesgo, debe ser examinada para determinar su adecuación o inadecuación, en la situación concreta en que se aplica al test.

• *Fuente de DIF no identificable*: Es el caso en que, examinados los ítemes, no se encuentra una variable que pueda explicar las diferencias de dificultad entre los grupos. En estas situaciones de *ausencia de explicación* se suelen atribuir los resultados estadísticos al error Tipo I, concluyendo que se ha rechazado una hipótesis nula que era verdadera y que no existe funcionamiento diferencial del ítem.

Si bien es cierto que los investigadores del tema recomiendan utilizar siempre que se pueda más de una medida del DIF en previsión de estos problemas, planteamos una *duda razonable* sobre esta supuesta solución. Si constatamos la ausencia de análisis de sesgo en las pruebas que se utilizan en nuestrro país,

parece más que razonable desconfiar de que las investigaciones que se hagan utilicen diversos métodos para prevenir las tasas de error Tipo I y Tipo II. Por esta razón, consideramos pertinente una reflexión crítica sobre este problema en previsión de estas situaciones más que probables de *supuesto* error Tipo I.

Atribuir a errores estadísticos el desacuerdo entre el análisis lógico y el análisis estadístico, parece una *solución de compromiso* sin una justificación aceptable. Si el análisis estadístico indica presencia de DIF y el análisis lógico no encuentra la variable que lo explica, ¿por qué suponer que el error está en el análisis estadístico, en lugar de aceptar como *medida de cautelar* la presencia de sesgo en el ítem? Teniendo en cuenta la dificultad que supone la identificación de la fuente de DIF, no parece descabellado dudar de la capacidad de identificación de la variable o variables responsables del funcionamiento diferencial en el ítem.

Englehard, Hansche, y Rutledge (1990) encontraron que el método de jueces también producía errores y numerosas ponencias del congreso anual del *National Council of Measurement in Education*, celebrado en abril de 1991 en Chicago (Breland, 1991; Bridgeman y Lewis, 1991; Mazzeo, Schmitt y Bleistein, 1991) muestran que determinar si los índices de DIF pueden ser atribuidos a factores relevantes o irrelevantes para el constructo, es bastante problemático y nada sencillo en muchos casos. Si se acepta, entonces, que las situaciones de discrepancia entre la estadística y los jueces son producto del error Tipo I, habrá que tener presente también el error Tipo II y suponer que algunos ítemes que no exhiben DIF posiblemente tienen un funcionamiento diferencial. Sin embargo, mientras se adoptan medidas de protección contra el error Tipo I, no se procede igual frente al error Tipo II que, en este caso, parece más grave. Es preferible equivocarse desechando un ítem que se sospeche sesgado y que no lo esté, a equivocarse en la dirección contraria. Si bien la posición *conservadora* suele ser siempre la más *segura* en estadística, en este caso tiene consecuencias sociales más graves: asume riesgos mayores que una posición más *liberal*.

Es ésta una controversia que emerge en alguna de las últimas publicaciones sobre el tema. Camilli y Shepard (1994) le dedican un epígrafe que no llega a dos páginas: pero expresan la necesidad de argumentar y explicar su posición *conservadora*, y reconocen que no es universalmente aceptada. Críticas como la de Davis Thissen, a la que Camilli y Shepard (1994) hacen referencia, deberán ser tenidas en cuenta cuando se plantee la relación entre sesgo y DIF. La perspectiva *liberal* de Thissen es partidaria de considerar que un ítem con DIF está sesgado hasta que no halla evidencia de lo contrario, y no al revés como se está planteando actualmente. La razón fundamental es que

> "[el] arte y ciencia de la psicología es demasiado impreciso para hacernos creíble que los especialistas de la medida procedente de grupos afectados, pueden emitir conclusiones seguras sobre las causas de la diferencial dificultad del ítem" (p. 150).

Por todo lo expuesto, la distinción entre *sesgo del ítem* y *funcionamiento diferencial*, ampliamente aceptada en la actualidad, vuelve a parecer como una diferenciación estratégica; pero no sustantiva. Si bien es verdad que permite delimitar problemas de conocimiento, no modifica la conceptualización del problema, y aunque tiene gran interés como procedimiento práctico hace depender de la capacidad explicativa del investigador, la decisión de que un ítem pueda ser considerado sesgado o simplemente con funcionamiento diferencial. De este modo, el sesgo no sería únicamente un problema del ítem, sino también de la capacidad explicativa del investigador. Parece más correcto, mantener los términos *potencialmente sesgado* y *sesgado* para diferenciar aquellos ítemes en los que no se dispone de una explicación de su DIF, de aquéllos en que se dispone ya de esta explicación. La consecuencias práctica de esta posición sería que los ítemes que exhiben funcionamiento diferencial sin que pueda identificarse claramente la fuente de este DIF, deberían ser eliminados por *potencialmente sesgados*, hasta que se pruebe mediante nuevos análisis que este resultado es, efectivamente, consecuencia del error Tipo I.

Otro problema no resuelto, común a todos los procedimientos de detección de sesgo interno, es su capacidad para detectar y eliminar el sesgo de los tests, puesto que el funcionamiento diferencial del ítem se obtiene a partir de la puntuación total en el test. Por tanto, si el test en su conjunto está sesgado, esa *porción* de sesgo común a todos los ítemes no será detectada por las técnicas DIF. Se produce un mecanismo de circularidad del que es imposible salir sin un criterio externo. Los métodos del sesgo interno, por definición, sólo pueden detectar discrepancias de sesgos. Es decir, en el caso en que todos los ítemes del test presenten un sesgo constante, los procedimientos del sesgo interno sólo podrán detectarlo en los ítemes, cuando exceda ese nivel de sesgo constante. Lord (1980) desarrolló un procedimiento, atribuido a Marco (1977), para *purificar* las escalas. En términos generales, este procedimiento se inicia identificando la presencia de DIF en los ítemes, y eliminando su contribución a la puntuación total de la escala, de modo que en el segundo paso se obtiene una estimación del nivel de los sujetos en el constructo a partir de los ítemes que no mostraron DIF en el primer análisis. Con esta nueva estimación, se estudia nuevamente la presencia de DIF en los ítemes. Este *procedimiento iterativo*, ha sido investigado por numerosos autores que han sugerido modificaciones diversas en la forma de llevarlo a cabo (Drasgow, 1987; Park y Lautenschlager, 1990; Kim y Cohen, 1992; Lautenschlager, Flaherty y Park, 1994), pudiendo resumir los hallazgos encontrados en la afirmación de Park y Lautenschlager (1990, p. 172): "los procedimientos no iterativos deben ser abandonados en los análisis de sesgo".

## 7. DEL FUNCIONAMIENTO DIFERENCIAL DE LOS ÍTEMES AL FUNCIONAMIENTO DIFERENCIAL DE LOS TESTS

Un nuevo enfoque sobre la pertinencia de considerar el funcionamiento diferencial de los ítemes de forma individualizada, ha despertado el interés por otra forma de abordar este problema, que ha pasado a conocerse como *Funcionamiento Diferencial de los Tests* o de acuerdo con las iniciales de su denominación en inglés, DTF.

La idea básica de este planteamiento es que si un test es construido como una unidad integrada por diversos grupos de ítemes, debería ser estudiado también como una unidad, y por tanto el *funcionamiento diferencial* debería abordarse en el conjunto del test, remitiendo a los conceptos de *Cancelación y Amplificación* de DIF (Wainer, Sireci y Thissen, 1991). Se puede hablar de *Cancelación de DIF* en un test cuando un conjunto de ítemes exhiben DIF contra el grupo Focal y otro conjunto de ítemes contra el grupo de Referencia, neutralizándose los efectos de ambos conjuntos. La cancelación se produce entonces, cuando el DIF es bidireccional y balanceado. La *Amplificación del DIF* se produce cuando un conjunto de ítemes que exhiben DIF unidireccional, actúan colectivamente para producir un efecto desfavorable en el grupo considerado, provocando así que este grupo tenga menor éxito en el test. Es posible que el análisis individualizado de los ítemes no permita detectar presencia de DIF en ellos, por no alcanzar un efecto suficientemente grande, y que sólo el análisis conjunto permita detectar los efectos acumulados en los distintos ítemes.

La consideración conjunta del test, fue planteada ya por algunos autores como Roznowski (1988), quien señaló que si las decisiones se toman a partir de las puntuación en el test, el DIF a un nivel del ítem, puede tener limitada importancia ya que pequeñas cantidades de DIF pueden cancelarse en el test, resultando éste un test perfectamente válido como unidad. Esto refuerza el planteamiento de Humphreys (1986), quien en forma permanente ha subrayado lo difícilmente realizable, y casi imposible, que resultaría pretender construir un test con ítemes unidimensionales. Por esta razón recomienda construir tests con ítemes de "rico contenido", y puesto que la multidimensionalidad causa DIF, controlar el DIF balanceando los ítemes de modo que se produzca la cancelación. Pero la tarea de balancear los ítemes debe hacerse además, con las debidas precauciones, ya que si por ejemplo, se colocaran todos los ítemes con DIF favorable para un grupo al final del test, este modo de balancear sería

desafortunado. En resumen, a juicio de este autor, balancear adecuadamente los ítemes que exhiben DIF, podría ser una forma satisfactoria de construir tests, y una estrategia más simple que pretender eliminar el DIF de los ítemes del test.

Wainer, Sireci y Thissen (1991) señalan algunas de las ventajas de considerar el funcionamiento diferencial del test, frente al estudio individualizado del funcionamiento diferencial de los ítemes. Estas ventajas, son precisamente la posibilidad de detectar la cancelación y la amplificación del DIF. Por un lado, si los ítemes con DIF están balanceados, el DTF detecta la cancelación; y por otro lado, el DTF permite descubrir DIF que por tener una cuantía pequeña queda enmascarado en los análisis estadísticos, y en cambio agregado el efecto de otros ítemes con DIF, producen una amplificación y determinan el funcionamiento desfavorable del test, para un grupo determinado. El poder de detección de funcionamiento diferencial, aumenta analizando el DTF , y este "aumento del poder estadístico de enfrentar el DIF nos aporta otro instrumento para asegurar la imparcialidad (del test)" (p. 199).

Wainer, Sireci y Thissen (1991) desarrollaron dos técnicas de detección de DTF basadas en el modelo TRI de ítemes politómicos. Aunque los ítemes dicotómicos han sido los más investigados, el desarrollo de técnicas para ítemes politómicos está recibiendo una atención creciente (Mellenbergh, 1995; Potenza y Dorans, 1995; Welch y Miller, 1995). Las técnicas de Wainer, Sireci y Thissen (1991) parten de la consideración de que la probabilidad de obtener una determinada puntuación en el test, es función de la habilidad latente ($\theta$) del sujeto. Mediante análisis previos, como análisis factorial, se puede determinar la presencia de diversas dimensiones en el test, y si se encuentra multidimensionalidad en él, se pueden obtener puntuaciones separadas para cada una de las dimensiones. El análisis de DTF se realiza estudiando el funcionamiento diferencial de cada una de las dimensiones, del mismo modo que se estudiaría el funcionamiento diferencial de un ítem cualquiera. Los dos métodos propuestos por estos autores, son una derivación de los métodos utilizados para el análisis de DIF: *método de criterio interno,* que valora la habilidad del sujeto a partir del propio test, incluyendo la dimensión estudiada, y *método de criterio externo,* que estima la habilidad a partir de instrumentos diferentes del test estudiado. La aplicación de esta metodología DTF para comparar hombres y mujeres en diversos tests de aptitudes, dio como resultado que, ítemes que parecían satisfactorios en los análisis de DIF, mostraron funcionamiento diferencial desfavorable a uno de los géneros, cuando se aplicó el análisis de DTF. Estos autores consideran que sus métodos, basados en modelos TRI, "pueden generalizarse al procedimiento de Mantel-Haenszel pudiendo aplicarse usualmente" (p.215).

Shealy y Stout (1993a, 1993b) también desarrollaron un procedimiento de detección de *sesgo simultáneo en los ítemes* (SIB, de acuerdo con las iniciales de su denominación en inglés: *Simultaneous Item Bias*), conocido como SIBTEST. La consideración de que diversos ítemes de un test *individualmente sesgados* pueden combinarse para producir una determinada puntuación en el test, genera la hipótesis de que esa puntuación tiene una carga de sesgo mayor. Esto es particularmente cierto, cuando el tamaño del DIF de los ítemes individuales no alcanza a ser lo suficientemente grande como para ser detectado por estasdísticas. Un ejemplo de este tipo podría ser el caso de un test de matemáticas formulado con un lenguaje sofisticado, cuya comprensión requiere un buen dominio de la lengua. En este caso, la combinación de la dificultad en matemáticas con la dificultad de comprensión de la lengua, puede producir sesgo del test contra poblaciones cuya primera lengua no es la del test, como ocurre con la gran población de hispanos en USA. Lo mismo se podría decir con respecto a poblaciones rurales, por ejemplo, si el lenguaje utilizado fuera típicamente urbano. El lenguaje sería en este caso un *determinante contaminador* que afectaría a la puntuación en el test. La distinción entre *habilidad objetivo* y *determinantes contaminadores* es un concepto básico en el procedimiento SIBTEST. La *habilidad objetivo* es el constructo que el test pretende medir, y los *determinantes contaminadores* serían los otros factores presentes en determinados ítemes, que tienen un efecto

sobre la puntuación en el test. Cuando estos *determinantes contaminadores* están balanceados en los ítemes del test, de modo que se produce un efecto de cancelación entre ellos, el test se considera insesgado, aunque los ítemes individuales exhiban DIF.

Mientras que el procedimiento de Wainer, Sireci y Thissen (1991) sigue un modelo TRI, el SIBTEST es paralelo al método de estandarización de Dorans y Kulick (1986), aunque fue desarrollado independientemente, como señalan Shealy y Stout (1993b). Esta es una ventaja que sus autores subrayan, en términos de economía de costes y de tiempo, que se añade a las ventajas expuestas por Nandakumar (1993), siendo las principales la posibilidad de detectar, además del DIF de los ítemes, su actuación simultánea y la posibilidad de diferenciar DTF del impacto debido a las diferencias entre los grupos en la habilidad que se pretende medir. Además, ofrece la posibilidad de comprender los mecanismos que determinan el funcionamiento diferencial entre los grupos, al considerar la habilidad en el constructo que se pretende medir o *habilidad objetivo*, en oposición a los *determinantes contaminadores* que saturan en mayor o menor medida los ítemes que exhiben DIF.

El SIBTEST es por tanto, un procedimiento no paramétrico diseñado para detectar DIF uniforme, y generalmente unidireccional. Para su ejecución, es necesario identificar los ítemes que saturan únicamente en la *habilidad objetivo*, a los que Shealy y Stout denominan *subset*, y a partir de la puntuación en este subset de ítemes unidimensionales, asignar a los sujetos a un determinado nivel en el constructo. La identificación de este subset de ítemes, deberá realizarse mediante análisis de multidimensionalidad, tales como el DIMTEST de Stout (1987) basado en TRI. Una vez determinados los niveles en el constructo que van a ser considerados, se comparan los sujetos de los grupos Focal y de Referencia, del mismo modo en que son comparados mediante cualquier procedimiento basado en tablas de contingencia. Este requerimiento es una limitación para su uso, si bien Shealy y Stout (1993b) proponen –en el caso en que no se disponga de un subset válido– aplicar la misma estrategia que en el procedimiento Mantel-Haenszel: estimar el nivel en el constructo, a partir de los ítemes que no hayan mostrado DIF en un examen preliminar, incluyendo el ítem objeto de estudio.

El procedimiento SIBTEST permitiría, además, diferenciar entre sesgo del test y DTF. Según Nandakumar (1993), si se utiliza una variable externa para determinar el nivel en el constructo, se estaría detectando sesgo, y cuando el criterio es interno se estaría detectando únicamente DTF. Esta distinción, teóricamente válida, adolece de los problemas ya señalados inherentes a cualquier pretensión de obtener una medida válida del nivel en el constructo: en primer lugar, la dificultad de encontrar otros instrumentos de medida pertinentes, y en segundo lugar, la dificultad aún mayor de garantizar que estos otros instrumentos no están igualmente afectados por sesgo o por funcionamiento diferencial. Sin embargo, la estrategia de obtener más de una medida de habilidad en el constructo puede reducir la tasa de error en la estimación. Mazor, Kanjee y Clauser (1995) encontraron que incluyendo más de una medida de habilidad en el análisis se mejoran las condiciones de estimación de DIF en los ítemes.

## 8. ALGUNAS FUENTES DE SESGO EN LOS TESTS

Aunque se han realizado algunos estudios a identificar estas fuentes potenciales de sesgo (Tittle, 1982; Reynolds y Brown, 1984; Scheuneman, 1984; 1985; Scheuneman y Gerritz, 1990), no se le ha dedicado la atención necesaria, y quizás la vuelta a la investigación con la técnica de jueces pueda ser una alternativa válida, como señalan Hambleton, Clauser, Mazor y Jones (1994).

Entre las posibles fuentes de sesgo contra grupos minoritarios, se han sugerido algunas con mayor o menor base empírica. Entre las causas de sesgo investigadas (Scheuneman, 1984), se pueden señalar las siguientes:

- *Contenido inapropiado*: Los tests se han construido principalmente a partir del vocabulario, conocimientos, y valores de la clase media occidental, por lo que otros grupos étnicos minoritarios, o de otros medios

sociales expuestos a estímulos diferentes, están en situación de desventaja al enfrentarse con un contenido en los tests que no les resulta familiar, o que les exige determinadas aptitudes para responder a los ítemes (Reynolds y Brown, 1984). Un problema muy generalizado es el de los cuestionarios de aptitudes que requieren habilidades de lectura, para responder a ítemes que pretenden medir constructos diferentes: razonamiento, cálculo, etc. (Scheuneman, 1984; Schmitt y Dorans, 1990; 1991).

• *Lenguaje inapropiado:* El lenguaje del test es un factor que afecta negativamente al grupo no familiarizado con él, como se ha comprobado en diversos estudios (Dorans y Kulick, 1983; Scheuneman, 1984; 1985). Los cuestionarios de aptitudes que requieren conocimientos de literatura, arte, música clásica, tienen mayor dificultad para sujetos de menor nivel socioeconómico (Jensen, 1980; Schmitt y Dorans, 1990; 1991; Mestre y Royer, 1991). Del mismo modo, ítemes que utilizan términos deportivos, presentan una dificultad mayor para mujeres que para hombres (Camilli y Shepard, 1994).

• *Formato inapropiado*: Características formales del test pueden ser también una fuente de sesgo. Por ejemplo, determinados formatos de respuesta han mostrado un efecto negativo sobre miembros de grupos minoritarios. Scheuneman (1985) realizó una exhaustiva investigación en el *Educational Testing Service*, en la que contrastó 16 hipótesis sobre causas posibles de sesgo en los ítemes: vocabulario utilizado, formato de respuesta de elección múltiple o cuantitativa, numeración romana o arábiga, palabras con sufijos o prefijos, una o varias respuestas falsas, una o varias respuestas verdaderas, lugar que ocupa la respuesta verdadera entre los distractores, utilización de símbolos o diagramas en la respuesta, etc. Comparando diferentes grupos étnicos obtuvo que muchos de estos formatos originaban un funcionamiento diferencial de los ítemes.

Uno de los formatos más investigados, que ha generado un método específico para la detección de sesgo, es el de aquellos ítemes en los que los sujetos deben elegir la respuesta correcta entre otras soluciones incorrectas o distractores. Si la relación que los sujetos establecen entre la pregunta y los distractores difieren en función de su grupo de pertenencia, el efecto puede ser la preferencia de uno de los grupos por un determinado distractor, que lleva a los miembros de ese grupo a cometer más errores. Esta preferencia diferencial puede deberse a que el contenido del distractor está más relacionado con la pregunta para el grupo que lo prefiere, colocándolo en situación de desventaja al añadir dificultad al ítem para sus miembros. Trabajos como los de Green, Crone, y Folk (1989), Schmitt y Dorans (1990, 1991), Dorans, Schmitt y Bleistein (1992), etc. han investigado este tipo de sesgo.

• *Muestras de estandarización inapropiadas:* Los grupos minoritarios no están suficientemente representados en las muestras con las cuales se estandariza el test. Este problema aparece incluso en los tests que gozan de gran *prestigio psicométrico*, como la escala de Wechsler. Wright e Isenstein (1977) ya criticaron el WISC-R porque se había estandarizado con una muestra de 2200 niños, entre los cuales sólo 330 pertenecían a grupos minoritarios. Recientemente, Maller (1994) encontró sesgo en los ítemes del WISCH-III, desfavorable a niños sordos frente a niños de audición normal. Van Dell (1994) encontró el mismo problema para niños de ámbito rural.

• *Condiciones de aplicación inapropiadas*: Las condiciones en que se aplica la prueba es otro factor de sesgo para sujetos no familiarizados con ellas. El efecto del examinador sobre los sujetos, sobre todo si se trata de niños, es uno de esos factores. Si éste pertenece al grupo mayoritario, sus características personales pueden ejercer un efecto intimidatorio que se refleja en una peor ejecución del test para miembros de otros grupos minoritarios (Reynolds y Brown, 1984). La insuficiente claridad y especificidad de las instrucciones, sobre todo en tareas nuevas, producen el mismo efecto (Scheuneman, 1984).

## CONCLUSIÓN

El problema del sesgo en los tests es una cuestión de gran importancia por las repercusiones sociales que lleva asociadas, importancia que debería reflejarse en decisiones más comprometidas acerca del uso indiscriminado

de tests. En este sentido, la investigación psicométrica debería reflejar la importancia del debate social sobre los problemas de discriminación que padecen diversos grupos y colectivos en nuestro medio, generando una dinámica de *corrección* y *prevención contra el sesgo* similar a la adoptada por el *Educational Testing Service* en USA.

El sesgo de género detectado en gran número de tests investigados en USA a partir del impacto de la crítica feminista, seguramente afecta a muchos de los que se utilizan en otros países, sin que este problema haya recibido la atención que merece. Si bien es cierto que despierta un interés creciente como *problema técnico,* se echa de menos una reflexión comprometida, sobre los supuestos e implícitos que los sustentan.

Es desde la revisión de la metodología que sustenta investigaciones sesgadas, desde donde se podrá desenmascarar el carácter ideológico de unas construcciones científicas que se pretenden neutras. Gran parte de las construcciones de la psicología diferencial, que muestran teorías todavía circulantes, e incluso revitalizadas en los últimos años, tienen como base empírica fundamental el uso de tests. Recientemente, se han vuelto a mostrar estudios que señalan la superioridad intelectual de los blancos, o la inferioridad de las mujeres: Instrumentos sesgados sólo pueden llevar a teorías sesgadas.

REFERENCIAS BIBLIOGRÁFICAS

AMEG Comission on Sex Bias in Measurement (1973). "AMEG Comission report on sex bias in interest measurement". *Measurement and Evaluation in Guidance,* 6, 171-177.

American Educational Research Association, American Psychological Association & National Council on Measurement in Education (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

American Psychological Association, American, Educational Research Association & National Council on Measurement in Education (1974). *Standards for educational and psychological tests.* Washington, DC: American Psychological Association.

American Psychological Association, Joint Committe on Testing Practices (1988). *Code of fair testing practices in education.* Washington, DC: Autor.

Angoff, W. H. (1982). "Use of difficulty and discrimination indices for detecting item bias". En: R. A. Berk (Ed.), *Handbook of methods for detecting tests bias* (pp. 96-116). Baltimore: Johns Hopkins University Press.

Angoff, W. H. (1993). "Perspectives on the theory and application of differential item functioning methodology". En: P. W. Holland y H. Wainer (Eds.), *Differential item functioning: Theory and practice* (pp. 3-23). Hillsdale: Erlbaum.

Berk, R. A. (1982). "Introduction". En: R. A. Berk (Ed.), *Handbook of methods for detecting test bias* (pp. 1-8). Baltimore: The Johns Hopkins University Press.

Bersoff, D. N. (1982). "The legal regulation of school psychology". En: C. R. Reynolds y T. B. Gutkin (Eds.) *The handbook of school psychology* (pp. 97-121). Nueva York: Wiley.

Binet, A. & Simon, T. (1916). *The development of intelligence in children.* Nueva York: Arno.

Breland, H. M. (1991, Abril). "The advanced Placement Test item format study" *Ponencia* presentada en el Congreso nacional del National Council on Measurement in Education, Chicago.

Bridgeman, B. & Lewis, C. (1991, Abril). "The predictive validity of Advanced Placement essay and multiple-choice scores". *Ponencia* presentada en el Congreso nacional del National Council on Measurement in Education, Chicago.

Camilli, G. (1993). "The case against DIF techniques based on internal criteria: Do item bias procedure obscure test fairness issues?". En: P. W. Holland y H. Wainer (Eds.), *Differential item functioning: Theory and practice* (pp. 397-417). Hillsdale: Lawrence Erlbaum.

Camilli, G. & Shepard, L. A. (1994). *Methods for Identifying Biased Test Items*. Thousand Oaks: Sage Publications.

Candell, G. L. & Hulin, C. L. (1987). "Cross-language and cross-cultural comparisons in scale translations: Independent sources of information about item nonequivalence". *Journal of Cross-Cultural Psychology,* 17, 417-440.

Cattell, R. B. (1940). "A culture-free intelligence test". *Journal of Educational Psychology,* 31, 161-179.

Cleary, T. A. (1968). "Test bias: Prediction of grades of Negro and white students in integrated colleges". *Journal of Educational Measurement,* 5, 115-124.

Cole, N. S. & Moss, P. A. (1989). "Bias in the test use". En Robert L. Linn (Ed.), *Educational Measurement* (pp. 201-219). New York: McMillan Publishing Company (3ª Edición).

Corulla, W. J. (1988). "A further psychometric investigation of the Sensation Seeking scale form V and its relationship to the EPQ-R and the I.7 impulsiveness Questionnaire". *Personality and Individual Differences,* 9, 277-287.

Corulla, W. J. (1989). "The relationship between the Strelau Temperament Inventory, sensation seeking and Eysenck's dimensional system of personality". *Personality and Individual Differences,* 10, 161-173.

Cruise, P. I. & Kimmel, E. W. (1990). *Changes in the SAT-Verbal: A study of trends in content and gender references, 1961-1987* (College Board Report nº 90-1). Nueva York: College Entrance Examination Board.

Darlington, R. B. (1971). "Another look at "cultural fairness"" *Journal of Educational Measurement,* 3, 71-82.

Darlington, R. B. (1978). "Cultural test bias: Comment on Hunter and Schmidt". *Psychological Bulletin,* 85, 673-674.

Delgado, C. (1994). "El sesgo sexual en la medición psicológica". *Revista de Psicología de El Salvador,* 51, 5-26.

Delgado, C. (1995). "Sesgo de género en la medición del neuroticismo". *Revista de Ciencias Sociales de la Universidad de Costa Rica,* 69, 51-66.

Delgado, C. (1997). "Psicología política y conciencia política de la psicología. Notas para una reflexión". En *Temas de Psicología* (V). Salamanca: Publicaciones Universidad Pontificia, p. 177-191.

Delgado, C. & Martín, M. F. (1997a). "Sesgo de género en la medición del Neuroticismo: Diferencias culturales". *Comunicación* presentada en el I Congreso Regional de Psicología para profesionales en América. Entrelazando la ciencia y la práctica en la psicología. México DF: 24 julio-1 agosto.

Delgado, C. & Martín, M. F. (1997b). "Aplicación de la técnica de Mantel-Haenszel a la detección de sesgo de género". *Actas* de VI Conferencia Española de Biometría, 141-143. Córdoba, 22-24 septiembre.

Donoghue, J. R. & Allen, N. L. (1993). "Thin versus thick matching in the Mantel-Haenszel procedure for detecting DIF". *Journal of Education Statistics,* 18, 131-154.

Dorans, N. J. & Holland, P. W. (1993). "DIF detection and description: Mantel-Haenszel and standarization". En P. W. Holland y H. Wainer (Eds.), *Differential*

item functioning: Theory and Practice (pp. 35-66). Hillsdale: Lawrence Erlbaum.

Dorans, N. J. & Kulick, E. M. (1983). *Assessing unexpected differential item performance of female candidates on SAT and TSWE forms administered in December 1977: An application of the standarization approach*. Princeton: Education Testing Service.

Dorans, N. J. & Kulick, E. M. (1986). "Demostrating the utility of the standarization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test". *Journal of Educational Measurement,* 23, 355-368.

Dorans, N. J.; Schmitt, A. P. & Blenstein, C. A. (1992). "The standarization approach to assessing comprehensive differential item functioning". *Journal of Educational Measurement,* 29, 309-319.

Drasgow, F. (1987). "Study of the measurement bias of two standardized psychological tests". *En Journal of Applied Psychology,* 72, 19-29.

Educational Testing Service (1987). *The ETS sensitivity review process: An overview*. Princeton: Educational Testing Service.

Eells, K; Davis, A.; Havighurst, R. J.; Herrick, V. E. & Tyler, R. W. (1951). *Intelligence and cultural differences*. Chicago: University of Chicago Press.

Ellett, F. S. (1980, Abril). "Fairness and the predictors". *Ponencia* presentada en el Congreso anual de la American Educational Research Association, Boston.

Ellis, B. R. (1989). "Differential Item Functioning: Implications for Test Translations". *Journal of Applied Psychology,* 74, 912, 921.

Ellwein, M. C.; Walsh, D. J.; Eads, G. M. & Miller, A. (1991). "Using readiness tests to route kindergarten students: The snarled intersection of psychometrics, policy, and practice". *Education Evaluation and Policy Analysis,* 13, 159-175.

Englehard, G.; Hansche, L. & Rutledge, K. E. (1990). "Accuracy of bias review judges in identifying differential item functioning on teacher certification tests". *Applied Measurement in Education,* 3, 347-360.

Francis, L. J. (1993). "The dual nature of the eysenckian neuroticism scales: A question of sex differences?". *Personality and Individual Differences,* 15, 43-59.

Green, D. R. (1975). "What does it mean to say a test is biased?". *Education and Urban Society,* 8, 33-52.

Green, B. F.; Crone, C. R. & Folk, V. G. (1989). "A method for studying Differential Distractor Functioning". *Journal of Educational Measurement,* 26, 147-160.

Hambleton, R. K.; Clauser, B. E.; Mazor, K. M. & Jones, R. W. (1993). "Advances in the detection of differentially functioning test items". *European Journal of Psychological Assessment,* 19, 1-18.

Hampson, S. E. (1988). *The construction of personality*. Londres: Routledge (2ª Edición).

Harmon, L. W. (1973). "Sexual bias in interest measurement". *Measurement and Evaluation in Guidance,* 5, 496-501.

Hilliard, A. G. (1979). "Standardization and cultural bias as impediments to the scientific study and validation of "intelligence"". *Journal of Research and Development in Education,* 12, 47-58.

Hilliard, A. G. (1984). "IQ Testing as the Emperor's New Clothes. A critique of Jensen's Bias in Mental Testing". En C. R. Reynolds y R. T. Brown (Eds.), *Perspectives on Bias in Mental Testing*. (pp. 139-169). Nueva York: Plenum Press.

Hofstee, K. B. (1990). "The use of everyday personality language for scientific purposes". *European Journal of Personality,* 4, 77-78.

Holland, P. W. & Thayer, D. T. (1988). "Differential item functioning and the Mantel-Haenszel procedure". En H. Wainer y H. I. Braun (Eds.) *Test validity* (pp. 129-145). Hillsdale: Lawrence Erlbaum.

Hulin, C. L. (1987). "A psychometric theory of evaluations of item scale translations: Fidelity across languages". *Journal of Cross-Cultural Psychology,* 18, 115-142

Humphreys, L. G. (1986). "An analysis and evaluation of test and item bias in the prediction context". *Journal of Applied Psychology,* 71, 327-333.

Hunter, J. E. & Schmidt, F. L. (1976). "A critical analysis of the statistical and ethical implications of various definitions of "tests bias"". *Phychological Bulletin,* 83, 1053-1071.

Hunter, J. E. & Schmidt, F. L. (1978). "Differential and single group validity of employement test by race: A critical analysis of three recent studies. *Journal of Applied Psychology,* 63, 1-11.

Jensen, A. R. (1969). "How much can we boost IQ and scholastic achievement?" *Harvard Educational Review,* 39, 1-123.

Jensen, A. R. (1980). *Bias in mental testing.* Nueva York: Free Press.

Kim, S.; Cohen, A. S. & Park, T. (1995). "Detection indifferential item functioning in multiple groups". *Journal of Educational Measurement,* 32, 261-276.

Kuhn, T. (1962). *The structure of scientific revolutions.* Chicago: University of Chicago Press.

Kuhn, T. (1970). *The Copernican Revolution.* Cambridge: Harvard University Press.

Larry, P. *et al. versus* Wilson Riles *et al.* (1979, Octubre). C71 2270 (United States District Court for the Northern District of California).

Lautenschager, G. J.; Flaherty, V. L. & Park, D. G. (1994). "ITR differential item functioning: an examination of ability scale purifications". *Educational and Psychological Measurement,* 54, 21-31.

Lewin, M. & Wild, C. L. (1991). "The impact of the feminist critique on tests, assessment, and methodology". *Psychology of Women Quarterly,* 15, 581-596.

Linn, R. L. (1973). "Fair test use in selection". *Review of Educational Research,* 43, 139-161.

Linn, R. L. (1979). "Issues of validity in measurement for competency-based programs". En M. A. Bunda y J. R. Sanders (Eds.), *Practice and problems in competency based measurement* (pp. 108-123). Washington: National Council on Measurment in Education.

Linn, R. L. (1980). "Issues of validity for criterion-referenced measures". *Applied Psychological Measurement,* 4, 547-561.

Linn, R. L. (1982). "Ability testing: Individual differences, prediction and differential prediction". En K. Wigdor y W. R. Garner (Eds.), *Ability testing: Uses, consequences, and controversies* (pp. 335-388). Washington: National Academy Press.

Linn, R. L. (1984). "Selection bias: Multiple meanings". *Journal of Education Measurement,* 21, 33-47.

Linn, R. L. & Harnisch, D. L. (1981). "Interactions between item content and group membership on achievement test items". *Journal of Education Measurement,* 18, 109-118.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale: Lawrence Erlbaum.

Maller, S. J. (1994, Noviembre). "Validity and item bias of the Wisc-III with deaf children". Universidad de Arizona: *Tesis* doctoral DAI-A 55/05, p. 1249.

Marco, G. L. (1977). "Item characteristic curve solutions to three intractable testing problems". *Journal of Educational Measurement,* 14, 139-160.

Mazor, K. M.; Clauser, B. E. & Hambleton, R. K. (1992). "The effect of sample size on the functioning of the Mantel-Haenszel statistic". *Educational and Psychological Measurement,* 52, 443-451.

Mazor, K. M.; Kanjee, A. & Clauser, B. E. (1995). "Using logistic regression and the Mantel-Haenszel with multiple ability estimates to detect differential item functioning". *Journal of Educational Measurement,* 32, 131-144.

Mazzeo, J.; Schmitt, A. P. & Bleistein, C. A. (1991, Abril). "Do women perform better, relative to men, on constructed-response rest or multiple-choice tests? Evidence from the Advanced Placement Examinations". *Ponencia* presentada en el congreso nacional de la National Council on Measurement in Education, Chicago.

Mc Nemar, Q. (1975). "On so-called test bias". *American Psychologist,* 30, 848-851.

Mellenbergh, G. J. (1995). "Conceptual notes on models for discrete polytomous item responses. Special Issue: Polytomous item response theory". *Applied Psychological Measurement,* 19, 91-100.

Mertz, W. R. (1974, abril). "A biased test may be fair, but what does that really mean?" *Ponencia* presentada en el Congreso de la California Educational Research Association, San Francisco.

Mestre, J.P. & Roger, J. M. (1991). "Cultural and linguist influences on Latino testing". En: G. D. Keller; J.R. Deneen y R. J. Magallan (Eds). *Assessment and Access: Hispanics in Higher Education.* New York: State University of New York Press. ps. 84-98.

Mischel, W. (1968). *Personality and Assessment.* New York: Wiley.

Montero, E. (1993). "Linguistic and cultural influences on differential item functioning for hispanic examinees in a standardized secondary level achievement test". *Tesis.* doctoral no publicada, Florida State University, Miami.

Muñiz, J. (1992). *Teoría clásica de los tests.* Madrid: Pirámide.

Nandakumar, R. (1993). "Simultaneous DIF amplification and cancellation: Shealy-Stout's test for DIF". *Journal of Educational Measurement,* 30, 293-311.

Osterlind, S. J. (1983). *Test Item Bias.* Beverly Hills: Sage Publications.

Park, D. G. & Lautenschlager, G. J. (1990). "Improving IRT Item Bias Detection with Iterative Linking and Ability Scale Purification". *Applied Psychological Measurement,* 14, 163-173.

Poortinga, Y. H. (1986). "Psychic unity versus cultural variation: An exploratory study of some basic personality variables in India and The Netherlands". *Informe* no publicado. Tilburg University.

Poortinga, Y. H. & Van de Vijver, F. J. R. (1987). "Explaining cross-cultural differences: Bias analysis and beyond". *Journal of Cross-Cultural Psychology,* 18, 259-282.

Potenza, M.T. & Dorans, N. J. (1995). "DIF assessment for polytomously scored items: A framework for classification and evaluation. Special Issue: Polytomous item response theory". *Applied Psychological Measurement,* 19, 23-37.

Reynolds, C.R. (1982a). "The problem of bias in psychological assessment". En: C.R.

Reynolds y T.B. Gutkin (Eds). *The handbook of school psychology* (ps. 178-208). New York: Wiley.

Reynolds, C. R. (1982b). "Methods for detecting construct and predictive bias". En R. A. Berk (Ed.), *Handbook of methods for detecting test bias* Baltimore: The Johns Hopkins University Press. ps. 200-227.

Reynolds, C. R. & Brown, R. T. (1984). "Bias in mental testing: An introduction to the issues". En: C. R. Reynolds y R. T. Brown (Eds.), *Perspectives on bias in Mental Testing* (pp. 1-39). Nueva York: Plenum Press.

Reynolds, C. R. & Paget, K. (1983). "National normative and reliability data for the Revised-Chidren's Manifiest Anxiety Scale". *School Psychology Review,* 12, 324-336.

Reynolds, C. R.; Plake, B. S. & Harding, R. D. (1983). "Item bias in the assessment of children's anxiety: Race and sex interaction on items of the Revised Children's Manifiest Anxiety Scale". *Journal of Psycho-Educational Assessment,* 1, 17-24.

Roznowoski, M. (1988). "Review of test validity". *Journal of Educational Measurement,* 25, 357-361.

Saville, P. & Blinkhorn, S. (1976). *Undergraduate personality by factored scales.* Windsor: NFER.

Scheuneman, J. D. (1979). "A method for assesing bias in test items". *Journal of Educational Measurement,* 16, 143-152.

Scheuneman, J. D. (1984). "A theoretical Framework for the Exploration of Causes and Effects of Bias in Testing". *Educational Psychologist,* 19, 219-225.

Scheuneman, J. D. (1985). "Exploration of causes of bias in test items". *Report GREB* Nº 81-21P. *Educational Testing Service, Report* 85-42.

Scheuneman, J. D. & Bleistein, C. A. (1989). "A consumer's guide to statistics for identifying differential item functioning". *Applied Measurement in Education,* 2, 255-275.

Scheuneman, J. D. & Gerritz, K. (1990). "Using Differential Item Functioning Procedures to Explor Sources of Item Difficulty and Group Performance Characteristics". *Journal of Educational Measurement,* 27, 109-131.

Schmitt, A. P. & Dorans, N. J. (1990). "Differential item functioning for minority examinees on the SAT". *Journal of Educational Measurement,* 27, 67-81.

Schmitt, A. P. & Dorans, N. J. (1991). "Factors related to differential item functioning for Hispanic examinees on the Scholastic Aptitude Test". En G. D. Keller, J. R. Deneen, y R. J. Magallan (Eds.), *Assessment and Access: hispanics in Higher Education* (pp. 12-37). Nueva York: Universidad del Estado de Nueva York.

Sharratt, S. (1993). *Feminismo y ciencia: Una relación problemática.* San José de Costa Rica: FLACSO.

Shealy, R. & Stout, W. F. (1993a). "An item response theory model for test bias". En P. W. Holland & H. Wainer (Eds.), *Differential item functioning: Theory and practice* (pp. 197-239). Hillsdale: Lawrence Erlbaum.

Shealy, R. & Stout, W. F. (1993b). "A model-based standarization approach that separates true bias/DIF from group ability differences and detects tests bias/DFT as well as item bias/DIF", *Psychometrika,* 58, 159-194.

Shepard, L. A. (1982). "Definitions of Bias". En R. A. Berk (Ed.) *Handbook of Methods for Detecting Test Bias* (pp. 9-30). Baltimore: The Johns Hopkins University Press.

Shepard, L. A. & Graue, M. E. (1993). "The moras of school readiness screening:

Research on test use and test validity". En B. Spodek (Ed.) *Handbook of research on the education of young children* (pp. 293-305). Nueva York: Macmillan.

Simon, A. & Thomas, A. (1983). "Means standard desviations and stability coefficients on the EPI for Further Education and College of Education students". *Personality and Individual Differences*, 4, 95-96.

Stern, W. (1914). *The psychological methods fro testing intelligence*. Baltimore: Warwick & York.

Stout, W. F. (1987). "A nonparametric approach for assessing latent trait unidimensionality". *Psychometrika*, 52, 589-617.

Thissen, D.; Steinberg, L. & Gerrard, M. (1986). "Beyond Group-Mean Differences: The Concept of Item Bias". *Psychological Bulletin*, 99, 118-128.

Thorndike, R. L. (1971). "Concepts of cullture-fairness". *Journal of Educational Measurement*, 8, 63-70.

Tittle, C. K (1982). "Use of judgmental methods in item bias studies". En R. A. Berk (Ed.) *Handbook of Methods for Detecting Test Bias* (pp. 31-63). Baltimore: The Johns Hopkins University Press.

Ukeje, I. C. (1990, Marzo). "Effective screening procedures for the identification of culturally differenty-educationally disadvantaged, intellectually gifted minority preschool children (educationally disadvantaged)". Universidad de New Jersey. *Tesis* doctoral DAI-A 51/09, p. 3024.

Van de Vijver, F. J. R. & Poortinga, Y. H. (1985). "A comment on McCauley and Colberg's conception of cross-cultural transportability of tests." *Journal of Educational Measurement* 22, 157-161.

Van de Vijver, F. J. R. & Poortinga, Y. H. (1991). "Testing Across Cultures." En R. K. Hambleton y J. Zaal (Eds.), *Advances in Educational Testing: Theory and Applications* (pp. 277-308). Boston: Kluwer Academic Publishers.

Van Dell, D. L. (1994, Enero). "An unusual variation in Wechsler Information subtest performance K-12". Universidad de Wyoming. *Tesis* doctoral DAI-B 55/07, p. 3001.

Wainer, H.; Sireci, S. G. & Thissen, D. (1991). "Differential Testlet Functioning: Definitions and detection". *Journal of Educational Measurement*, 28, 197-219.

Welch, C. J. & Miller, T. R. (1995). "Assessing differential item functioning in direct writing assessments: Problems and an example". *Journal of Educational Measurement*, 32, 163-178.

Williams, R. L. (1970). "Danger: Testing deshumanizing black children". *Clinical Child Psychology Newsletter*, 9, 5-6.

Wright, B. D. & Isenstein, V. R. (1977). *Psychological tests and minorities*. Rockville, NIMH DHEW Publication.

Yang, K. & Bond, M. H. (1990). "Explorating implicit personality theories with indigenous or imported constructs". *Journal of Personality and Social Psychology*, 58, 1087-1095.

*Carmen Delgado A.*
*Departemento de Psicología*
*Universidad Pontificia de Salamanca*
*c/Compañía 5 37002 Salamanca, España.*

*Fax (923) 26 24 56*
*E- Mail: cdelgado@gugu.usal.es*