Metodología de las Ciencias del Comportamiento 2(2),151-165 ©aemcco, 2000

Equiparación de puntuaciones: exigencias actuales y retos de cara al futuro

María José Navas[1]

UNED

Resumen

La equiparación de puntuaciones es un proceso fundamental cuando se trabaja con distintos instrumentos de medida, ya que representa el medio básico de que se dispone para poder garantizar una comparación adecuada de las puntuaciones obtenidas en las distintas pruebas. A continuación se describen muy brevemente las líneas de trabajo y los desarrollos más importantes que han tenido lugar en el campo durante las dos últimas décadas. Seguidamente se examina la situación actual de los tests con el fin de ver cuál es el papel que juega en la actualidad la equiparación y en qué situaciones y por qué surge la necesidad de equiparar. Por último, se reflexiona acerca de los retos que el presente plantea al futuro.
PALABRAS CLAVE: *Equiparación de puntuaciones*

Abstract

EQUATING SCORES: STATE OF THE ART AND CHALLENGES TO THE FUTURE. Equating scores is a fundamental process when working with different tests, since it represents the basic means available to guarantee an appropriate comparison of the scores obtained in different tests. The latest research and the most important developments that have taken place in the field during the last two decades are briefly described. Next the state of the art of the tests is examined to see what part equating plays at the present time, and in what situations and why equating is necessary. Finally, some challenges for the future are outlined.
KEY WORDS: *Equating scores.*

La equiparación de puntuaciones es un proceso fundamental cuando se trabaja con distintos instrumentos de medida, ya que representa el medio básico de que se dispone para poder garantizar una comparación adecuada de las puntuaciones obtenidas en las distintas pruebas: equiparar consiste simplemente en derivar puntuaciones equivalentes para poder comparar las puntuaciones obtenidas en distintos tests que, obviamente, deben medir el mismo constructo o característica. Éste no es el único requisito que han de cumplir dos tests para poder equiparar sus puntuaciones sino que se han de satisfacer, además, los tres siguientes: invarianza en la población, simetría y equidad (Angoff, 1984).

Son muchas las situaciones en las que se requiere la aplicación frecuente de distintas formas de una misma prueba o test. Por ejemplo, en un examen de oposición en el que se convoca a los aspirantes para distintas fechas resulta extremadamente conveniente disponer de formas alternativas de la prueba de examen, por razones estrictamente de seguridad. También es necesario disponer de distintas formas de un test cuando se desea medir en repetidas ocasiones a un mismo individuo o colectivo con el fin de evaluar, por ejemplo, su progreso académico o un posible cambio en sus actitudes. En cualquiera de estos casos, para poder comparar las puntuaciones obtenidas en las distintas formas del test es necesario ponerlas previamente en la misma escala. Ése es justamente el cometido del proceso de equiparación de las

[1]Dirección postal para correspondencia: María José Navas. Departamento de Metodología de las Ciencias del Comportamiento. Facultad de Psicología. UNED. Ciudad Universitaria, s/n. 28040 MADRID (Spain).

puntuaciones: garantizar una comparación adecuada derivando puntuaciones equivalentes para las pruebas implicadas en el proceso, hallar la transformación que permite establecer una métrica común entre las puntuaciones de los tests. Esto solo es posible si se dan alguna de las siguientes condiciones: (1) los tests tienen ítems en común, (2) las muestras de sujetos a las que se aplican los tests están solapadas o (3) se trabaja con muestras aleatorias extraídas de la misma población de sujetos.

El abanico de métodos disponibles para derivar puntuaciones comparables es bastante grande (véase tabla 1). Por un lado, están los métodos clásicos -el lineal y el equipercentil- y, por otro, los métodos basados en la teoría de respuesta al ítem. Muy brevemente, con el método lineal se consideran puntuaciones equivalentes las que corresponden a idénticas puntuaciones típicas y con el método equipercentil las que tienen centiles iguales. La teoría de respuesta al ítem ofrece la posibilidad de equiparar distintos tipos de puntuaciones (puntuaciones θ, observadas, verdaderas) y aporta metodología propia para la equiparación. Para una revisión de estos métodos, véase el trabajo de Kolen y Brennan (1995) en inglés o el de Navas (1996) en castellano.

Tabla 1. Clasificación de los métodos de equiparación

<table>
<tr><th></th><th></th><th>Método Lineal</th><th>Método Equipercentil</th></tr>
<tr><td rowspan="2">Métodos Clásicos</td><td>Basados en el grupo sintético</td><td>Tucker
Levine para tests con la misma fiabilidad
Levine para tests con distinta fiabilidad
Método de Braun y Holland (1982)
Estimación de frecuencias</td><td>Estimación de frecuencias</td></tr>
<tr><td>Basados en los datos</td><td>Anclaje externo:
. Lord (1955)
. Doble equiparación
. Anclaje predictor
. Anclaje predicho
Anclaje interno:
. Lord (1955)
. Thurstone (1925)
. Swineford y Fan (1957)</td><td>Directo
Anclaje predictor
Anclaje predicho</td></tr>
<tr><td rowspan="3">Métodos basados en la TRI</td><td>Basados en los momentos</td><td colspan="2">M. estándar de la media y la desviación típica
M. de la media y la desviación típica robustas
M. iterativo de la media y la desviación típica robustas y ponderadas</td></tr>
<tr><td>Basados en la curva característica</td><td colspan="2">M. de la curva característica del test
M. del χ^2 mínimo</td></tr>
<tr><td>Otros métodos</td><td colspan="2">M. de las b's fijas
M. de calibración concurrente</td></tr>
</table>

No es fácil decidirse por un método u otro, a pesar de que buena parte de la literatura sobre el tema se ocupa justamente de esta cuestión. 'La comparación de los métodos clásicos y los basados en la TRI está sustancialmente influida por muchos factores, como la fiabilidad de los tests que se van a equiparar, las propiedades de los tests de anclaje, el nivel de habilidad de las muestras y el tipo de tests a equiparar. Estos factores pueden producir tanta o más variación en los resultados de la equiparación que la simple elección entre métodos (Skaggs, 1990, p.105). A título meramente orientativo, se recomienda utilizar el método lineal cuando el tamaño de la(s) muestra(s) es pequeño, los tests a equiparar no son muy diferentes y solo se

requiere una gran precisión en una zona de la escala que no se encuentra muy alejada de la media (por ejemplo, cuando el test se utiliza para la certificación académica y el punto de corte está próximo a la media). Se recomienda utilizar el método equipercentil y los métodos basados en la teoría de respuesta al ítem cuando el tamaño muestral es grande y se requiere precisión a lo largo de toda la escala. En cualquier caso, el constructor o el usuario del test ha de tener muy presente que, sea cual sea el método elegido, su funcionamiento se va alejando progresivamente del óptimo conforme más diferentes son los grupos de sujetos a los que se administran los tests y más distintas son las formas del test administradas, así como cuanto más difieren entre sí las especificaciones para los ítems comunes y para los ítems únicos de las pruebas. Kolen y Brennan (1995, pp. 269-271) ofrecen en una tabla una relación de las situaciones para las que es más apropiado cada uno de los métodos.

Acerca del pasado

Durante mucho tiempo, la equiparación ha recibido muy poca atención en la literatura científica. De hecho, hasta la década de los 80 prácticamente la única publicación -al menos de impacto- es el capítulo escrito por Angoff *Scales, Norms and Equivalent Scores* en la edición de 1971 del libro *Educational Measurement*, publicado posteriormente por el *Educational Testing Service* en 1984 como documento independiente y recientemente incluido en el libro editado por Ward, Stoker y Murray-Ward (1996).

La semilla de la equiparación comienza a germinar realmente cuando crece la demanda de evaluación, que hay que vincular al movimiento de *accountability* que surge con fuerza en la década de los setenta en los EE.UU. La razón es que se había hecho una considerable inversión económica en educación y se deseaba conocer los resultados de la misma. Esto se tradujo en un notable incremento en el número y variedad de programas de evaluación que utilizaban formas múltiples de un test (Brennan, 1987) y supuso el detonante para una línea de trabajo e investigación que con el tiempo se ha revelado ciertamente fructífera.

La década de los 80 es testigo de la aparición en la literatura de un número creciente de trabajos sobre el tema (Cook, Dunbar y Eignor, 1981; Divgi, 1981; Jaeger, 1981; Kolen, 1981; Lord y Wingersky, 1984; Petersen, Cook y Stocking, 1983; Stocking y Lord, 1983; Vale, 1986). En 1982 Holland y Rubin editan un libro que se va a convertir en un texto clásico de referencia. En este libro se recogen discusiones generales de algunas cuestiones, los métodos de equiparación basados en la teoría de respuesta al ítem y la metodología utilizada en el *Educational Testing Service* para poner las puntuaciones de distintos tests en una escala común. Asimismo, son ya varios los textos en los que se presenta el tema con tratamiento de capítulo independiente (Crocker y Algina, 1986; Goldstein, 1986; Hambleton y Swaminathan, 1985; Lord, 1980; Thorndike, 1982). La tercera edición del libro *Educational Measurement* se publica en 1989 e incluye de nuevo un magnífico capítulo sobre equiparación, esta vez escrito por Petersen, Kolen y Hoover. De gran interés son también los artículos publicados por Cook y Eignor (1983, 1989), Kolen (1988) y Skaggs y Lissitz (1986). En estos trabajos se puede encontrar un buen tratamiento del tema, examinándose con cierto detalle cuestiones fundamentales para la equiparación. La tercera edición de los *Standards for Educational and Psychological Testing* (AERA, APA y NCME, 1985) es la primera edición que contempla explícitamente este tema, al que dedica cuatro estándares (del 4.6 al 4.9); sin embargo, en la cuarta edición (AERA, APA y NCME,

1999) el número de estándares relativos a la equiparación es ya el doble que en la anterior (del 4.10 al 4.17). En 1987 la revista *Applied Psychological Measurement* le dedica un número monográfico a esta cuestión. En 1989, en la reunión anual de la *American Educational Research Association* se celebró un *symposium* bajo el título *Selecting samples for equating: To match or not to match.* Las comunicaciones presentadas en este *symposium* fueron publicadas como un número monográfico en la revista *Applied Measurement in Education* (primer número de 1990).

En la década de los 90 la tónica sigue siendo la misma. Se publica el segundo libro sobre equiparación (Kolen y Brennan, 1995), un auténtico tratado sobre el tema, y los manuales de psicometría publicados en castellano comienzan también a abordar esta cuestión (Martínez Arias, 1995; Santisteban, 1990), algunos de ellos incluso con tratamiento de capítulo independiente (Muñiz, 1996, 1998). Asimismo, el volumen de investigación dedicado en esta década a la equiparación es bastante notable. Continúan los trabajos destinados a comparar distintos métodos de equiparación (Baker y Al-Karni, 1991; Han, Kolen y Pohlmann, 1997; Kim y Cohen, 1996; Kolen y Harris, 1990; Yen y Burket, 1997) y se proponen alternativas y se refinan algunos de estos métodos (Liou y Cheng, 1995a; Liou, Cheng y Wu, 1999; Wang y Kolen, 1996; Zeng y Kolen, 1995). Hay también algunos trabajos ocupados en formular el error típico de algunos métodos de equiparación (Hanson, Zeng y Kolen, 1993; Liou y Cheng, 1995b; Liou, Cheng y Johnson, 1997; Zeng, 1993; Zeng y Cope, 1995; Zeng, Hanson y Kolen, 1994). Otros se centran en las distintas técnicas de suavizado de las distribuciones de las puntuaciones, con el fin de eliminar en la medida de lo posible las fluctuaciones muestrales de la equiparación equipercentil (Kolen, 1991; Little y Rubin, 1994; Livingston, 1993; Zeng, 1995). Otros trabajos se ocupan de la equiparación vertical, esto es, de la equiparación de las puntuaciones de tests con distinto nivel de dificultad, que es algo más compleja que la equiparación horizontal, en la que los tests tienen aproximadamente la misma dificultad (Camilli, 1999; Camilli, Yamamoto y Wang, 1993; de Gruijter y de Jong, 1991; Harris, 1991; Schulz, Perlman, Rice y Wright, 1992; Yamamoto y Mazzeo, 1992; Yen y Burket, 1997). Otra línea de investigación trata de obtener puntuaciones comparables en pruebas aplicadas en proyectos de evaluación diferentes, llevados a cabo a nivel de estado o de distrito escolar, a nivel nacional (*National Assessment of Educational Progress, Voluntary National Test*), internacional (*Third International Mathematics and Science Study*), etc. (Beaton y González, 1993; Bloxom, Pashley, Nicewander y Yan, 1995; Ercikan, 1997; Linn, 1993; Linn y Kiplinger, 1995; Pashley y Phillips, 1993; Williams, Rosa, McLeod, Thissen y Sanford, 1998). También se han dedicado algunos esfuerzos a los problemas específicos que se pueden plantear en la equiparación de tests referidos al criterio (Norcini, 1990; Norcini y Shea, 1992; Norcini, Shea y Grosso, 1991; Norcini, Shea y Lipner, 1994) y en los tests que constan de ítems de respuesta abierta y de respuesta construida (Huynh y Ferrara, 1994; Loyd, Engelhard y Crocker, 1995; Wainer, Wang y Thissen, 1994; Wang, Wainer y Thissen, 1995; Yen y Ferrara, 1997). Finalmente, algunos trabajos comienzan a apuntar hacia la utilización de modelos de teoría de respuesta al ítem que trabajan con ítems no dicotómicos sino politómicos. Así, el modelo que ha recibido más atención hasta el momento es el modelo de respuesta graduada de Samejima (Baker, 1992, 1997; Cohen y Kim, 1998; Kim y Cohen, 1995); algo menos el modelo nominal de Bock (Baker, 1993, 1997) y el modelo de crédito parcial (Huynh y Ferrara, 1994).

En torno al presente

En la actualidad, es incuestionable que los tests forman parte del entramado social en el que transcurre la vida del hombre, utilizándose en muchas ocasiones para tomar decisiones que pueden ser importantes en la vida de las personas. Normalmente, desde los primeros años de escolarización, los niños comienzan a responder a tests que servirán para medir sus aptitudes intelectuales y su potencial para el aprendizaje; se utilizarán también para certificar su progreso y su nivel de conocimientos, así como para orientarlos vocacionalmente en el futuro. Más tarde, los tests pueden abrir (o cerrar) las puertas a la enseñanza universitaria y, una vez en el mercado laboral, ayudarán a decidir quién deberá cubrir un determinado puesto de trabajo y quién será promocionado; además, los tests desempeñarán también un papel importante a la hora de determinar si un sujeto está o no bien adaptado a su entorno, si dispone de recursos suficientes para resolver los problemas que se le plantean en su vida cotidiana o si, por el contrario, necesita algún tipo de terapia para reforzar alguno de estos aspectos. El consumo de tests en nuestra sociedad es grande: son muchos los millones de escolares, trabajadores, opositores y militares que en la actualidad son examinados regularmente con tests.

Sin embargo, a pesar de ser muy utilizados los tests han sido también muy criticados; Ebel y Frisbie (1986) y Hopkins (1998) se refieren a esto como a la paradoja de los tests. Los tests han recibido críticas de todo tipo y pelaje aunque se les ha acusado, sobre todo, de ser herramientas reaccionarias al servicio del poder, de constituir barreras para la igualdad social y de oportunidades económicas, simplemente porque los tests han servido para revelar diferencias entre grupos, y el tema de las diferencias entre grupos es muy espinoso, levanta muchas ampollas. Ahora bien, estas críticas no han traído consigo una disminución en el uso de los tests -ahí está la paradoja, tanto las críticas como el uso de los tests han sido una constante a lo largo de las últimas décadas- pero sí han tenido un impacto importante en la forma de trabajar con los tests. Por ejemplo, no hace muchos años en EE.UU. se consideraba innecesario informar al sujeto de su puntuación en una prueba de acceso a la Universidad. En la actualidad, no solo es preciso informarle de su puntuación sino que es obligatorio, además, mostrar públicamente el test y su plantilla de corrección. Lo cierto es que la continua crítica, el intenso debate y la discusión social que este sistema generalizado de tests ha generado ha llevado a que en EE.UU. la presión que ejerce la opinión pública y la misma legislación hayan obligado a que se hagan públicas las pruebas y sus resultados. Las consecuencias de esta decisión están lejos de ser triviales. Por un lado, ha conducido a una visibilidad creciente de las cuestiones técnicas relativas a la equiparación, dado que los constructores de tests se han visto obligados a justificar y explicar públicamente sus métodos de equiparación. Es decir, el tema de la equiparación ha pasado del reducto constituido por organismos dedicados a la evaluación y especialistas en el campo de la medida a la arena pública, al escrutinio directo de educadores, legisladores y del público en general (Holmes, 1986). Por otro lado, este sistema generalizado de tests sometido al escrutinio público conduce a que los tests apenas si pueden ser reutilizados, se 'queman' muy rápidamente y solamente se pueden volver a aplicar como mucho algunos de sus items, pero no la prueba como tal, que es preciso mostrar públicamente. Por consiguiente, se asiste a una necesidad creciente de construir formas alternativas de un mismo test cada vez con mayor frecuencia.

Desde esta perspectiva los bancos de ítems constituyen herramientas de inestimable valor a la hora de construir tests (distintas pruebas, formas alternativas, paralelas, etc.). Pues bien, la construcción de un banco de items -y especialmente la incorporación de nuevos items al banco y la actualización de las estimaciones de sus parámetros- supone inevitablemente el recurso a procesos de equiparación. Para empezar, es difícil pensar en construir un banco de items sin planear cuidadosamente un diseño de equiparación que posibilite poner en una misma escala las estimaciones de los parámetros de los distintos items aplicados a grupos diferentes de sujetos, pero las necesidades de equiparación continúan estando más presentes si cabe en las fases posteriores de gestión y mantenimiento del banco. El hecho de trabajar con bancos de ítems permite además organizar las pruebas en una serie de secciones operativas (secciones que se van a tener en cuenta para obtener la puntuación total del sujeto y que, por tanto, será necesario dar a conocer sus ítems o, al menos, ponerlos a disposición de los sujetos examinados) y de secciones variables (secciones que no serán contabilizadas en la puntuación total del sujeto). La ventaja fundamental que presenta esta configuración de la prueba en secciones operativas y variables es que, gracias a que no es necesario difundir los ítems de las secciones variables, se pueden incorporar nuevos ítems calibrados al banco y es posible mantener escalas de referencia, garantizado que en todo momento se informa en la misma escala de medida y se cumple con la normativa vigente. En este sentido, son de enorme interés los diseños de preequiparación de sección y del ítem (*véase* Petersen y col., 1989).

La necesidad de equiparar surge también en otras situaciones de gran interés. Como se acaba de indicar, los tests han sido acusados de ser instrumentos de control o de represión social al servicio de la clase económica o políticamente dominante, por lo que se ha examinado con lupa el posible sesgo y/o funcionamiento diferencial de los ítems de los tests en distintos grupos definidos por variables tales como el género, la raza, la cultura, etc. -véase en este monográfico el trabajo de Hidalgo y López-Pina (2000)-. Pues bien, la equiparación de puntuaciones suele ser también necesaria cuando se investiga el posible sesgo o funcionamiento diferencial de los items de un test, especialmente cuando el estudio se aborda desde la teoría de respuesta al ítem, ya que antes de poder comparar las curvas características de los items obtenidas en los distintos grupos es condición *sine qua non* que las estimaciones de los parámetros estén en la misma escala. En la actualidad, hay toda una línea de investigación que ha puesto claramente de manifiesto que la introducción de métodos iterativos a la hora de equiparar las puntuaciones mejora considerablemente la tasa de detección de items sesgados e insesgados (Candell y Drasgow, 1988; Kim y Cohen, 1992; Lautenschlager, Flaherty y Park, 1994; Navas y Gómez, 1994; Park y Lautenschlager, 1990). En realidad, lo anterior es válido no solo cuando se estudia el funcionamiento diferencial de los ítems de un test desde la óptica de la teoría de respuesta al ítem sino siempre que se utilice esta teoría de tests y se disponga de dos conjuntos distintos de estimaciones de los parámetros (de los ítems o de los sujetos). La razón tiene que ver con el hecho de que la escala θ, aunque invariante, es arbitraria, esto es, no está determinado ni su origen ni su unidad de medida, por lo que esos dos conjuntos de estimaciones no están realmente en la misma escala, si bien presentan valores linealmente relacionados. Por consiguiente, en el marco de la teoría de respuesta al ítem es preciso determinar las constantes de esa trans-

formación lineal que permite poner en la misma métrica las estimaciones obtenidas para los parámetros en dos ocasiones distintas.

Sobre el futuro

No son pocos los retos que el presente plantea al futuro. Antes al contrario, son muchos los desafíos que se presentan a los tests para poder responder a las crecientes demandas que se generan en la sociedad -cada vez más exigentes y diversificadas- y a la necesidad de una mayor claridad y transparencia a lo largo de todas las fases implicadas en la utilización de los tests, equiparación incluida. Parece claro que en el futuro la práctica profesional con tests deberá ser extremadamente cauta y prudente y tendrá que caracterizarse por una transparencia todavía mayor en sus métodos y formas de operar. Ésta es, sin duda, una buena vía para acallar críticas y evitar los procesos judiciales en los que se han visto inmersos los tests en las últimas décadas.

Antes de ver cuáles son las demandas o exigencias específicas que los tests del futuro plantean en relación a la equiparación de sus puntuaciones, es preciso apuntar una cuestión que no ha recibido suficiente atención en la literatura científica. Ésta tiene que ver con la fase de cierre del proceso de equiparación: la evaluación de su calidad, esto es, la determinación de si la conversión derivada da lugar a puntuaciones efectivamente equivalentes. A pesar de la indudable importancia de esta fase, ésta es todavía poco frecuente en muchos estudios de equiparación y son muy pocos los trabajos que han tratado de sistematizar esta cuestión. Dos excepciones las constituyen los trabajos de Brennan y Kolen (1987) y Harris y Crouse (1993). Como ya se indicó anteriormente, hay algunos trabajos dedicados a formular y cuantificar el error aleatorio que se introduce en el proceso obteniendo el correspondiente error típico de equiparación, pero se echan de menos trabajos más relacionados con el componente sistemático del error, donde se opere con distintos métodos, en contextos diferentes de equiparación y haciendo variar distintos factores; para ello, se puede utilizar el paradigma circular o encadenado (se equipara un test consigo mismo, bien directamente, bien a través de una cadena o eslabones de tests) o recurrir a la simulación. En esta línea, los trabajos de Baker (1996, 1997, 1998) constituyen una notable excepción.

La investigación y los avances en esta fase son decisivos ya que pueden facilitar enormente la tarea -nada fácil- de decidir qué método utilizar para equiparar o incluso si equiparar o no. En efecto, hay situaciones en las que equiparar introduce más error del que eliminaría (Angoff y Schrader, 1981; Dorans y Lawrence, 1990; Hanson, 1992; Harris, 1991; Kolen y Harris, 1990), por lo que se dispone de herramientas estadísticas que ayudan a decidir si es o no necesario llevar a cabo una equiparación (Hanson, 1992; Kolen y Jarjoura, 1987). Una vez establecida la conveniencia de equiparar hay que enfrentarse al problema de decidir qué método se va a utilizar. También se han desarrollado algunas ayudas estadísticas para decidir entre distintos métodos clásicos (Budescu, 1987; Jaeger, 1981; Zeng, 1995); en ocasiones se utiliza como criterio para optar por uno u otro método los resultados ofrecidos por la literatura en pruebas similares o se escoge el método que muestra los resultados más consistentes con resultados anteriores obtenidos con esa prueba. Sin embargo, es evidente que es la investigación sobre la equiparación con cualquiera de los paradigmas disponibles la que debe decir la última palabra sobre esta importante cuestión.

En relación a los tests del futuro, éstos parecen apuntar hacia el diseño de formas menos restrictivas de ítems y hacia la construcción de nuevas formas de tests, tanto en su contenido como en su formato y modo de administración, si bien en muchos aspectos el futuro ya está aquí.

En efecto, en los últimos años el foco de atención de investigadores, educadores y legisladores se ha ido desplazando hacia la denominada medición auténtica (*authentic measurement, performance assessment*), que pone el énfasis en evaluar no tanto habilidades básicas sino habilidades cognitivas de orden superior (es decir, habilidades integradoras, metacognitivas, estrategias de solución de problemas, etc. -véase el trabajo de Prieto y Delgado (2000) en este monográfico-. La evaluación de este tipo de habilidades supone un reto para los tests, ya que implica el recurso a pruebas con ítems de respuesta abierta y, especialmente, a ítems de respuesta construida que pueden proporcionar un conocimiento más profundo y un análisis más detallado de niveles superiores en la actuación de los sujetos que las pruebas de elección múltiple. Ahora bien, la equiparación de este tipo de pruebas reviste una complejidad mayor que la de los clásicos tests de elección múltiple. Así, la puntuación supone el recurso a jueces o calificadores y, por tanto, el error suele ser mayor que en las pruebas de elección múltiple; no resulta fácil en muchos casos utilizar los diseños más habituales de equiparación y, lo que es más importante, dado que el número de ítems en estas pruebas suele ser bastante pequeño, no se puede garantizar que se muestrea adecuadamente el dominio de interés con las pruebas, el recorrido de la escala puede ser bastante pequeño y resulta difícil obtener estimaciones estables de θ, lo que hace cuestionable la utilización de métodos basados en la teoría de respuesta al ítem. La comunidad investigadora tendrá que hacer, por tanto, un esfuerzo importante para dar respuesta a estas nuevas demandas de evaluación y a los problemas técnicos que éstas traen consigo. Al mismo tiempo, se requiere un esfuerzo paralelo en el desarrollo de métodos de equiparación basados en modelos para respuestas politómicas. En la década de los noventa ya hay trabajos que apuntan tímidamente en estas direcciones, si bien esta línea de trabajo debe consolidarse en el futuro.

Una cuestión que no puede ser pasada por alto es el hecho de que las pruebas de respuesta construida demandan del sujeto respuestas múltiples o, en cualquier caso, complejas y elaboradas, muy difíciles de resumir en las tradicionales escalas unidimensionales que, aunque tremendamente útiles, proporcionan una representación -en ocasiones- algo parcial de la realidad. La realidad es multidimensional y más cuando lo que se pretende evaluar son habilidades cognitivas de orden superior, conocimientos conceptuales complejos o procesos cognitivos. La multidimensionalidad es un hecho que no se puede obviar y que hay que afrontar con las herramientas metodológicas adecuadas. Ahora bien, ¿cómo equiparar las puntuaciones de pruebas que son decididamente multidimensionales? Lo que se suele hacer es utilizar los procedimientos habituales -unidimensionales- de equiparación. Hay una serie de trabajos que tratan de determinar en qué medida las puntuaciones obtenidas tras aplicar estos procedimientos se pueden considerar puntuaciones equivalentes cuando éstas están midiendo más de una dimensión, esto es, cuál es el impacto de la multidimensionalidad en los procedimientos unidimensionales de equiparación (Bolt, 1999; Camilli, Wang y Fesq, 1995; de Champlain, 1996; Dorans y Kingston, 1985). Davey, Oshima y Lee (1996) han dado un paso más y han extendido y adaptado los procedimientos unidimensionales de equiparación para su utilización con

modelos multidimensionales de teoría de respuesta al ítem; incluso han elaborado software para ello (Lee y Oshima, 1996). Un trabajo pionero en esta línea es el de Hirsch (1989).

Respecto al formato y modo de administración de las pruebas, el tremendo impulso experimentado por la tecnología del ordenador en las últimas décadas es en buena parte responsable de la cada vez más generalizada aplicación informatizada de las pruebas (incluso a través de Internet) y de las pruebas adaptativas (véase el trabajo de Ponsoda, Olea y Revuelta en este mismo número). En relación a estas últimas, hay algunos trabajos que reclaman un sistema de puntuación para este tipo de prueba que sea más fácil de entender y mucho más familiar para los usuarios de los tests que la métrica θ (van der Linden, 1999; Stocking, 1996). En particular, se propone utilizar como puntuación el número de aciertos, eso sí, corrigiendo la diferencia -intencional- en la dificultad de los ítems administrados en las pruebas adaptativas mediante la correspondiente equiparación. Eignor (1993) reflexiona sobre las dificultades que se plantean al derivar puntuaciones equivalentes obtenidas en pruebas adaptativas y en pruebas de lápiz y papel.

Por último, una línea de trabajo ya perfilada en la década de los noventa y que parece destinada a consolidarse es la equiparación de las puntuaciones obtenidas al aplicar distintos tests en proyectos de evaluación diferentes. De hecho, en las tres últimas reuniones anuales de la *American Educational Research Association/ National Council on Measurement in Education* (1998-2000), la mayor parte de los *simposia* organizados sobre equiparación se dedican justamente a esta cuestión. De lo que se trata, en definitiva, es de ver si es posible comparar los resultados obtenidos en el marco de una evaluación realizada dentro de un estado con los resultados obtenidos en una evaluación nacional como el *National Assessment of Educational Progress* o en el marco de un estudio internacional: se trata de predecir la distribución de las puntuaciones en un test a partir de las puntuaciones obtenidas en otro test aplicado en un ámbito distinto (estatal, nacional, internacional). Sin lugar a dudas, éste constituye un procedimiento tremendamente económico e informativo. Sin embargo, los resultados obtenidos hasta el momento ponen de manifiesto que de momento conviene ser cauto.

Esta línea de trabajo ilustra una tendencia que se observa de forma bastante clara en los trabajos realizados en los últimos años y es la tendencia a relajar los requisitos que en principio hay que satisfacer para derivar puntuaciones realmente equivalentes, para así poder equiparar puntuaciones de pruebas que, por ejemplo, no tienen la misma fiabilidad o que incluso no miden exactamente el mismo constructo o característica. De este modo, se hace cada vez más necesario utilizar la distinción terminológica entre equiparación y escalamiento para lograr comparabilidad (AERA, APA y NCME, 1985; Linn, 1993; Mislevy, 1992), reservado el término *equiparación* para la derivación de puntuaciones estrictamente intercambiables en tests construidos con las mismas especificaciones para medir el mismo constructo y donde la ecuación o tabla de equivalencia cumple los requisitos de simetría e invarianza en la población. Cualquier otra situación que requiera poner en la misma métrica las puntuaciones de dos tests debería ser etiquetada como *escalamiento para lograr comparabilidad*, esto es, como un proceso que permite obtener puntuaciones en la misma escala, puntuaciones comparables pero no totalmente intercambiables, es decir, que no da lo mismo aplicar un test u otro a un sujeto de cualquier nivel de habilidad. Los procedimientos empleados en la equiparación y en el escalamiento

para lograr comparabilidad son los mismos, no así el significado de las puntuaciones derivadas con uno y otro. Lo cierto es que la utilización cada vez más frecuente de tests y los cambios que en los últimos años ha experimentado este campo están obligando a los psicómetras a reconsiderar y ampliar el horizonte de la equiparación, planteándose nuevas formas de comparar y ajustar las puntuaciones obtenidas en distintas pruebas (Kolen y Brennan, 1995). Sirva como ilustración de lo anterior el hecho de que en la tercera edición de los *Standards for Educational and Psychological Testing* el capítulo dedicado, entre otras, a estas cuestiones se denominaba *Scaling, Norming, Score Comparability, and Equating*; en la cuarta edición el término *equating* simplemente ha desaparecido del título del capítulo.

Referencias

AERA, APA y NCME (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

AERA, APA y NCME (1999). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Angoff, W. H. (1971). Scales, norms and equivalent scores. En R. L. Thorndike (Ed.), *Educational Measurement.* Washington, DC: American Council on Education.

Angoff, W. H. (1984). *Scales, norms and equivalent scores.* Princeton, NJ: Educational Testing Service.

Angoff, W. H. y Schrader, W. B. (1981). *A study of alternative methods for equating rights scores to formula scores* (RR-81-8). Princeton, NJ: Educational Testing Service.

Baker, F. B. (1992). Equating tests under the graded response model. *Applied Psychological Measurement, 16,* 87-96.

Baker, F. B. (1993). Equating tests under the nominal response model. *Applied Psychological Measurement, 17(3),* 239-251.

Baker, F. B. (1996). An investigation of the sampling distributions of equating coefficients. *Applied Psychological Measurement, 20(1),* 45-57.

Baker, F. B. (1997). Empirical sampling distributions of equating coefficients for graded and nominal response instruments. *Applied Psychological Measurement, 21(2),* 157-172.

Baker, F. B. (1998). An investigation of the item parameter recovery characteristics of a Gibbs sampling procedure. *Applied Psychological Measurement, 22(2),* 153-169.

Baker, F. B. y Al-Karni, A. (1991). A comparison of two procedures for computing IRT equating coefficients. *Journal of Educational Measurement, 28,* 147-162.

Beaton, A. E. y Gonzalez, E. J. (1993). *Comparing the NAEP Trial State Assessment results with the IAEP International results* (Report prepared for the National Academy of Education Panel on the NAEP Trial State Assessment). Stanford, CA: National Academy of Education.

Bloxom, B., Pashley, P., Nicewander, W. A. y Yan, D. (1995). Linking to a large scale assessment: An empirical evaluation. *Journal of Educational and Behavioral Statistics, 20,* 1-26.

Bolt, D. M. (1999). Evaluating the effects of multidimensionality on IRT true score equating. *Applied Measurement in Education, 12(4),* 383-408.

Brennan, R. L. (1987). Introduction to problems, perspectives and practical issues in equating. *Applied Psychological Measurement, 11(3),* 221-224.

Brennan, R. L. y Kolen, M. J. (1987). Some practical issues in equating. *Applied Psychological Measurement, 11(3),* 279-290.

Braun, H. I. y Holland, P. W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. En P. W. Holland y D.B. Rubin (Eds.), *Test Equating.* New York: Academic Press.

Budescu, D. V. (1987). Selecting an equating method: Linear or equipercentile- *Journal of Educational Statistics, 12(1),* 33-43.

Camilli, G. (1999). Measurement error, multidimensionality, and scale shrinkage: A reply to Yen and Burket. *Journal of Educational Measurement, 36(1),* 73-89.

Camilli, G., Wang, M. y Fesq, J. (1995). The effects of dimensionality on equating the Law School Admissions Test. *Journal of Educational Measurement, 32,* 79-96.

Camilli, G., Yamamoto, K. y Wang, M. (1993). Scale shrinkage in vertical equating. *Applied Psychological Measurement, 17, 4,* 379-388.

Candell, G. L. y Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied Psychological Measurement, 12(3),* 253-260.

Champlain, A. de (1996). The effect of multidimensionality on IRT true-score equating for subgroups of examinees. *Journal of Educational Measurement, 33,* 79-96.

Cohen, A. S. y Kim, S. K. (1998). An investigation of linking methods under the graded response model. *Applied Psychological Measurement, 22(2),* 116-130.

Cook, L. L., Dunbar, S. A. y Eignor, D. R. (1981). *IRT equating: A flexible alternative to conventional methods for solving practical testing problems.* Comunicación presentada en la reunión anual de la AERA/NCME, Los Angeles.

Cook, L. L. y Eignor, D. R. (1983). Practical considerations regarding the use of item response theory to equate tests. En R.K. Hambleton (Ed.), *Applications of item response theory.* Vancouver, BC: Educational Reseach Institute of British Columbia.

Cook, L. L. y Eignor, D. R. (1989). Using item response theory in test score equating. En R.K. Hambleton (Ed.), Applications of item response theory. *International Journal of Educational Research, 13, 2,* 161-173.

Crocker, L. y Algina, J. (1986). *Introduction to classical and modern test theory.* Nueva York: Holt, Rinehart and Winston.

Davey, T., Oshima, T. C. y Lee, K. (1996). Linking multidimensional item calibrations. *Applied Psychological Measurement, 20(4),* 405-416.

Divgi, D. R. (1981). Model-free evaluation of equating and scaling. *Applied Psychological Measurement, 5,* 203-208.

Dorans, N. J. y Kingston, N. M. (1985). The effects of violations of unidimensionality on the estimation of item and ability parameters and on IRT equating of the GRE verbal scale. *Journal of Educational Measurement, 22(4),* 249-262.

Dorans, N. J. y Lawrence, I. M. (1990). Checking the statistical equivalence of nearly identical test editions. *Applied Measurement in Education, 3,* 245-254.

Ebel, R. L. y Frisbie, D. A. (1986). *Measuring educational achievement.* Englewood Cliffs, NJ: Prentice-Hall.

Eignor, D. R. (1993). *Deriving comparable scores for computer adaptive and conventional tests: An example using the SAT* (RR 93-55). Princeton, NJ: Educational Testing Service.

Ercikan, K. (1997). Linking Statewide Tests to National Assessment of Educational Progress: Accuracy of combining test results across states. *Applied Measurement in Education, 10(2),* 145-159.

Goldstein, H. (1986). Models for equating test scores and for studying the comparability of public examinations. En D. L. Nuttall (Ed.), *Assessing educational achievement.* Lewes, Sussex: Falmer Press.

Gruijter, D. N. M. de y Jong, J. H. A. L. de (1991). Item-rest regressions, item response functions and the relation between test forms. *Applied Psychological Measurement, 15(1),* 25-34.

Hambleton, R. K. y Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston: Kluwer-Nijhoff.

Han, T., Kolen, M. y Pohlman, J. (1997). A comparison among IRT true- and observed-score equatings and traditional equipercentile equating. *Applied Measurement in Education, 10(2),* 105-122.

Hanson, B. A. (1992). *Testing for differences in test score distributions using log-linear models.* Comunicación presentada en la reunión anual de la AERA, San Francisco.

Hanson, B. A., Zeng, L. y Kolen, M. J. (1993). Standard errors of Levine linear equating. *Applied Psychological Measurement, 17(3),* 225-237.

Harris, D. J. (1991). A comparison of Angoff's Design I and Design II for vertical equating using traditional and IRT methodology. *Journal of Educational Measurement, 28(3),* 221-235.

Harris, D. J. y Crouse, J. D. (1993). A study of criteria used in equating. *Applied Measurement in Education, 6(3),* 195-240.

Hidalgo, M.D. y López-Pina, J.A. (2000). Funcionamiento diferencial de los ítems: presente y perspectivas de futuro. *Metodología de las Ciencias del Comportamiento, 2(2),* 167-182.

Hirsch, T. M. (1989). Multidimensional equating. *Journal of Educational Measurement, 26, 4,* 337-349.

Holland, P. W. y Rubin, D. B. (Eds.) (1982). *Test equating.* Nueva York: Academic Press.

Holmes, S. E. (1986). Test equating and credentialing examinations. Special issue: Testing concerns in credentialing health professionals. *Evaluation and the Health Professions, 9(2),* 230-249.

Hopkins, K. D. (1998). *Educational and psychological measurement and evaluation.* Boston: Allyn and Bacon.

Huynh, H. y Ferrara, S. (1994). A comparison of equal percentile and partial credit equatings for performance-based assessments composed of free-response items. *Journal of Educational Measurement, 31(2),* 125-141.

Jaeger, R. M. (1981). Some exploratory indices for selection of a test equating method. *Journal of Educational Measurement, 18,* 23-38.

Kim, S. y Cohen, A. S. (1992). Effects of linking methods on detection of DIF. *Journal of Educational Measurement, 29(1),* 51-66.

Kim, S. K. y Cohen, A. S. (1995). A minimum chi-2 method for equating tests under the graded response model. *Applied Psychological Measurement, 19(2),* 167-176.

Kim, S. K. y Cohen, A. S. (1996). *A comparison of linking and concurrent calibration under item response theory.* Comunicación presentada en la reunión anual de la AERA, Nueva York.

Kolen, M. J. (1981). Comparison of traditional and item response theory methods for equating tests. *Journal of Educational Measurement, 18(1),* 1-11.

Kolen, M. J. (1988). Traditional equating methodology. *Educational Measurement: Issues and Practice, 7(4),* 29-36.

Kolen, M. J. (1991). Smoothing methods for estimating test score distributions. *Journal of Educational Measurement, 28(3),* 257-282.

Kolen, M.J. y Brennan, R.L. (1995). *Test equating: Methods and Practices.* Nueva York: Springer-Verlag.

Kolen, M. J. y Harris, D. J. (1990). Comparison of item preequating and random groups equating using IRT and equipercentile methods. *Journal of Educational Measurement, 27(1),* 27-39.

Kolen, M. J. y Jarjoura, D. (1987). Analytic smoothing for equipercentile equating under the common item nonequivalent populations design. *Psychometrika, 52(1),* 43-59.

Lautenschlager, G. J., Flaherty, V. L. y Park, D. G. (1994). IRT differential item functioning: An examination of ability scale purifications. *Applied Psychological Measurement, 54(1),* 21-31.

Lee, K. y Oshima, T. C. (1996). *IPLink: Multidimensional and unidimensional IRT linking* [Computer program]. Atlanta: Georgia State University.

Linden, W. J. van der (1999). *Adaptive testing with equated number-correct scoring* (RR 99-02). Enschede: Twente University.

Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education, 6,* 83-102.

Linn, R. L. (1996). Linking assessments. En M. B. Kane y R. Mitchell, *Implementing performance assessment. Promises, problems, and challenges.* Mahwah, NJ: LEA.

Linn, R. L. y Kiplinger, V. L. (1995). Linking statewide tests to the National Assessment of Educational Progress: Stability of results. *Applied Measurement in Education, 8(2),* 135-155.

Liou, M. y Cheng, P. E. (1995a). Equipercentile equating via data-imputation techniques. *Psychometrika, 60(1),* 119-136.

Liou, M. y Cheng, P. E. (1995b). Asymptotic standard error of equipercentile equating. *Journal of Educational and Behavioral Statistics, 20(3),* 259-286.

Liou, M., Cheng, P. E. y Johnson, E. G. (1997). Standard errors of the kernel equating methods under the common-item design. *Applied Psychological Measurement, 21(4),* 349-369.

Liou, M., Cheng, P. E. y Wu, C. (1999). Using repeaters for estimating comparable scores. *British Journal of Mathematical and Statistical Psychology, 52,* 273-284.

Little, R. J. A. y Rubin, D. B. (1994). Test equating from biased samples, with application to the Armed Services Vocational Aptitude Battery. *Journal of Educational and Behavioral Statistics, 19(4),* 309-336.

Livingston, S. A. (1993). Small-sample equating with log-linear smoothing. *Journal of Educational Measurement, 30(1),* 23-40.

Lord, F. M. (1955). Equating test scores. A maximum likelihood solution. *Psychometrika, 20,* 193-200.

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

Lord, F. M. y Wingersky, M. (1984). Comparison of IRT true-score and equipercentile observed-score equatings. *Applied Psychological Measurement, 8,* 452-461.

Loyd, B., Engelhard, G. y Crocker, L. (1995). Achieving form-to-form comparability: Fundamental issues and proposed strategies for equating performance assessments of teachers. *Educational Assessment, 3(1),* 99-110.

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

Mislevy, R. J. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects.* Princeton, NJ: Educational Testing Service.

Muñiz, J. (1996). *Psicometría.* Madrid: Universitas.

Muñiz, J. (1998). *Teoría clásica de los tests.* Madrid: Pirámide.

Navas, M. J. (1996). Equiparación. En J. Muñiz (Coord.), *Psicometría.* Madrid: Universitas.

Navas, M. J. y Gómez, J. (1994). *Comparison of several bias detection techniques.* Comunicación presentada en el 23rd International Congress of Applied Psychology, Madrid.

Norcini, J. J. (1990). Equivalent pass/fail decisions. *Journal of Educational Measurement, 27,* 59-66.

Norcini, J. J. y Shea, J. (1992). The reproducibility of standards over groups and occasions. *Applied Measurement in Education, 5(1),* 63-72.

Norcini, J. J., Shea, J. y Grosso, L. J. (1991). The effect of number of experts and common items on cutting score equivalents based on expert judgment. *Applied Psychological Measurement, 15,* 241-246.

Norcini, J., Shea, J. y Lipner, R. (1994). The effect of anchor item characteristics on equivalent cutting scores. *Applied Measurement in Education, 7(3),* 187-194.

Park, D. G. y Lautenschlager, G. J. (1990). Improving IRT item bias detection with iterative linking and ability scale purification. *Applied Psychological Measurement, 14,* 163-173.

Pashley, P. y Phillips, G. W. (1993). *Toward world-class standards.* Princeton, NJ: ETS.

Petersen, N. S., Cook, L. L. y Stocking, M. S. (1983). IRT versus conventional equating methods: A comparative study of scale stability. *Journal of Educational Statistics, 8,* 137-156.

Petersen, N. S., Kolen, M. J. y Hoover, H. D. (1989). Scaling, norming and equating. En R. L.Linn (Ed.), *Educational measurement.* Nueva York: Macmillan.

Prieto, G. y Delgado, A.R. (2000). Utilidad y representación en la psicometría actual. *Metodología de las Ciencias del Comportamiento, 2(2),* 111-128.

Santisteban, C. (1990). *Psicometría. Teoría y práctica en la construcción de tests.* Madrid: Norma.

Schulz, E. M., Perlman, C., Rice, W. K. y Wright, B. D. (1992). Vertically equating reading tests: An example from Chicago Public Schools. En M.Wilson (Ed.), *Objective measurement: Theory into Practice.* Norwood, NJ: Ablex Publishing Corporation.

Skaggs, G. (1990). To match or not to match samples on ability for equating: A discussion of five articles. *Applied Measurement in Education, 3(1),* 105-113.

Skaggs, G. y Lissitz, R. W. (1986) . IRT test equating: Relevant issues and a review of recent research. *Review of Educational Research, 56,* 495-529.

Stocking, M. L. (1996). An alternative method for scoring adaptive tests. *Journal of Educational and Behavioral Statistics, 21(4),* 365-389.

Stocking, M. y Lord, F. M. (1983). Developing a common metric in IRT. *Applied Psychological Measurement, 7(2),* 201-210.

Swineford, F. y Fan, C. (1957). A method of score conversion through item statistics. *Psychometrika, 22,* 185-188.

Thorndike, R. L. (1982). *Applied psychometrics.* Boston: Houghton Mifflin Company.

Thurstone, L. L. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology, 16,* 433-449.

Vale, C. D. (1986). Linking item parameters onto a common scale. *Applied Psychological Measurement, 10,* 333-344.

Wainer, H., Wang, X. y Thissen, D. (1994). How well can we compare scores on test forms that are constructed by examinees' choice- *Journal of Educational Measurement, 31(3),* 183-199.

Wang, T. y Kolen, M. J. (1996). A quadratic curve equating method to equate the first three moments in equipercentile equating. *Applied Psychological Measurement, 20(1),* 27-43.

Wang, X., Wainer, H. y Thiseen, D. (1995). On the viability of some untestable assumptions in equating exams that allow examinee choice. *Applied Measurement in Education, 8(3),* 211-225.

Ward, A. W., Stoker, H. W. y Murray-Ward, M. (1996). *Educational measurement :Origins, theories and explications.* Lanham: University Press of America.

Williams, V. S. L., Rosa, K. R., McLeod, L. D., Thissen, D. y Sanford, E. E. (1998). Projecting to the NAEP scale: Results from the North Carolina End-Of-Grade Testing Program. *Journal of Educational Measurement, 35(4),* 277-296.

Yamamoto, K. y Mazzeo, J. (1992). Item response theory scale linking in NAEP. *Journal of Educational Statistics, 17(2),* 155-173.

Yen, W. M. y Burket, G. R. (1997). Comparison of item response theory and Thurstone methods of vertical scaling. *Journal of Educational Measurement, 34(4),* 293-314.

Yen, W. M. y Ferrara, S. (1997). The Maryland School Performance Assessment Program: Performance assessment with psychometric quality suitable for high stakes usage. *Educational and Psychological Measurement, 57(1),* 60-84.

Zeng, L. (1993). A numerical approach for computing standard errors of linear equating. *Applied Psychological Measurement, 17(2),* 177-186.

Zeng, L. (1995). The optimal degree of smoothing in equipercentile equating with postsmoothing. *Applied Psychological Measurement, 19(2),* 177-190.

Zeng, L. y Cope, R. T. (1995). Standard error of linear equating for the conterbalanced design. *Journal of Educational and Behavioral Statistics, 20(4),* 337-348.

Zeng, L. y Kolen, M. J. (1995). An alternative approach for IRT observed-score equating of number-correct scores. *Applied Psychological Measurement, 19(3),* 231-240.

Zeng, L., Hanson, B. A. y Kolen, M. J. (1994). Standards errors of a chain of linear equating. *Applied Psychological Measurement, 18(4),* 369-378.