

Didáctica

LA EQUIPARACIÓN DE PUNTUACIONES EN PROCESOS DE COMPARACIÓN DE PRUEBAS DIFERENTES

Juan Sebastián Pacheco–Villamil*
Universidad Nacional de Colombia, Colombia

No es difícil darse cuenta de que el uso de pruebas e instrumentos psicológicos en la actualidad presenta un importante auge debido a su utilización en procesos masivos de selección y admisión en muchos contextos. Estos instrumentos de medición psicológica se hacen indispensables si se tiene en cuenta que además de intercambios y homologaciones, la demanda de cupos, becas, y trabajos, excede ampliamente la oferta que se presenta en las sociedades de hoy. Muchas instituciones, organizaciones y entidades se sirven de estas pruebas como herramientas para seleccionar a los individuos con las mejores cualidades, aptitudes o conocimientos. Pero su utilidad no beneficia solamente a quienes seleccionan, puesto que adicionalmente se garantiza la imparcialidad en los procesos de evaluación, razón por la cual, la utilización de estas “pruebas objetivas” ha pasado a tomar un carácter legal en procesos estatales de selección y evaluación como medio para brindar transparencia e igualdad de oportunidad entre candidatos o aspirantes. Como resultado de este auge hemos presenciado un continuo proceso de consolidación de la utilización de pruebas para la evaluación que a su vez ha ocasionado una gran proliferación de nuevos instrumentos de medición psicológica aplicados a toda clase de públicos.

Al utilizar estos instrumentos de medición para tomar decisiones dentro de procesos de selección, promoción, orientación, certificación, etc., razones de seguridad pueden exigir la aplicación de diferentes versiones de prueba en cada ocasión para mantener la confidencialidad y la seguridad, puesto que, si las mismas preguntas son usadas en cada aplicación de la prueba, las personas que toman el examen en una segunda oportunidad tendrán obvias ventajas sobre quienes la toman por primera vez (Holland & Rubin, 1982).

De esta manera, el panorama que nos ofrecen las condiciones actuales de evaluación hace que la comparabilidad o equiparación de puntuaciones sea 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 distintas pruebas (Navas, 1996). Se advierte con esto que la construcción de escalas comunes de comparación de puntajes tiene una enorme utilidad en la práctica cuando se trata de tomar decisiones sobre asignación de cupos o empleos con base en pruebas diferentes de un mismo atributo, o en procesos que impliquen comparar el desempeño de dos candidatos que han presentado pruebas diferentes.

* E-mail: jspachecov@unal.edu.co

Garantizar la comparabilidad de los resultados en pruebas diferentes obtenidos por un mismo individuo o grupo ha sido un tema de larga trayectoria en la psicometría y de interés creciente en los últimos cinco años. En este sentido resulta interesante subrayar que de una ausencia casi total o un somero tratamiento del tema en los libros clásicos de psicometría, se ha pasado a un gran volumen de producción al respecto desde diferentes perspectivas, de hecho en los últimos años se ha observado un importante incremento en el número de publicaciones por parte de revistas científicas especializadas (Herrera, 2005). A pesar de lo anterior, por lo menos en Colombia, aún se presenta un vacío a nivel conceptual que se manifiesta en el desconocimiento del término, de sus acepciones, y de sus implicaciones prácticas, además se presenta una carencia de trabajos que revisen o evalúen los avances de la comparación de pruebas, avances que en nuestro país permitirían optimizar la utilización de la información proporcionada en los procesos masivos de evaluación.

El concepto de equiparación

En su comienzo, la comparación de puntuaciones de pruebas diferentes centró su trabajo en el emparejamiento de las puntuaciones de formas paralelas de prueba, sin embargo, la expansión en el desarrollo y utilización de los instrumentos de medición psicológica ha llevado a la necesidad de comparar puntuaciones de una amplia variedad de pruebas que, en la mayoría de los casos, no cuentan con formas alternas o paralelas. Las formas paralelas de prueba son versiones de una prueba que se construyen con contenidos y especificaciones psicométricas similares. Estadísticamente hablando, las formas alternas de prueba son consideradas equivalentes o paralelas cuando características psicométricas como sus medias, varianzas y niveles de confiabilidad son iguales para la población de examinados que toman la prueba.

De acuerdo con Navas (1996), Muñiz (1997) y Kolen (2004) la mayoría de revisiones citan a Angoff (1984) y a Lord (1980) como los pioneros en este tema. Estos autores entienden la equiparación como un proceso encaminado al desarrollo de un sistema de conversión de las unidades de una prueba a las unidades de otra, de manera que sus resultados sean comparables o equivalentes. El problema consiste así en establecer una correspondencia entre las puntuaciones comparadas con el fin de hallar una métrica común para dos o más medidas de un mismo atributo, de manera que se puedan comparar adecuadamente los resultados de individuos diferentes o de instrumentos diferentes que midan el mismo atributo.

A partir de la misma concepción del término, Skaggs & Lissitz (1986) afirman que para efectos de continuidad, se necesita un sistema de conversión que permita colocar en una escala común las puntuaciones obtenidas en versiones recientes de una prueba con las versiones más antiguas de la misma, razón por la cual la investigación sobre la equiparación adquiere gran importancia para las organizaciones encargadas del desarrollo y estandarización de pruebas. Así lo anotan Holland & Rubin (1982) cuando mencionan que la equiparación implica los procedimientos a los cuales recurre el *Educational Testing Service ETS* en Estados Unidos para garantizar que los supuestos implícitos (puntajes intercambiables) son correctos.

Adicionalmente, Skaggs & Lissitz (1986) plantean dos tipos principales de equiparación: Horizontal y vertical. En la “*equiparación horizontal*” se cuenta con formas de prueba construidas para medir el mismo constructo con el mismo nivel de dificultad, para la misma

población de examinados. En la “*equiparación vertical*”, las pruebas son diseñadas para evaluar un mismo constructo pero con diferentes niveles de dificultad, siendo aquí el propósito la equiparación de pruebas que miden la misma habilidad pero en un amplio rango de dificultad. Desde las perspectivas teórica y práctica la equiparación vertical es mucho más complicada que la equiparación horizontal.

Reconocer que la equiparación pretende lograr puntuaciones totalmente intercambiables de modo que resulte indiferente la prueba que tome el examinado, hace que la similitud de los constructos evaluados en cada una de las pruebas repercuta en el nivel de comparación que puede lograrse. Una medida es intercambiable con otra medida si ambas estiman el mismo atributo y se expresan en la misma métrica. Sin embargo, los supuestos necesarios para un proceso de equiparación frecuentemente no se cumplen debido a que la mayoría de las pruebas no se diseñan para ser comparadas. Factores como este evidencian que los procesos de equiparación pueden verse alterados por diversas circunstancias. Kolen (2001) proporciona varios ejemplos donde demuestra que el tipo de comparación de puntuaciones menos claro ocurre precisamente en los procesos de equiparación. El problema más común tiene que ver con que las puntuaciones comparadas se interpreten como intercambiables cuando en realidad no lo son, lo cual podría afectar negativamente a algunos examinados debido a la elaboración de inferencias inadecuadas sobre la ejecución esperada en la prueba que no han tomado, a partir de su ejecución en otra prueba.

Cuando los métodos de la equiparación son aplicados a pruebas que difieren en contenido, dificultad o confiabilidad, las puntuaciones resultantes no pueden usarse como si se tratase de puntuaciones intercambiables (Yin, Brennan & Kolen, 2004). Según autores como Dorans & Holland (2000) la igualdad de la fiabilidad sería una condición necesaria pero no suficiente para la equiparación. En este sentido Pommerich, Harris, Hanson & Sconing (2004) sostienen que para que la comparación sea considerada como una equiparación deben sostenerse condiciones exigentes. Así, para poder derivar puntuaciones realmente equivalentes es necesario satisfacer los siguientes requisitos:

1. Todas las pruebas deben medir el mismo constructo con la misma fiabilidad.
2. La ecuación o tabla de equivalencia obtenida debe ser independiente del grupo de sujetos utilizados para derivarla. Este requisito se conoce como el de la *invarianza en la población*, que por cierto es muy difícil de satisfacer en la práctica por lo que autores como Petersen, Cook & Stocking (1983) señalan la conveniencia de dar indicaciones sobre las características de las muestras utilizadas para llevar a cabo la equiparación.
3. La equiparación debe ser simétrica, es decir, de modo que el resultado sea el mismo independientemente de la prueba que se tome como base a la hora de hacer equivalentes las puntuaciones en las pruebas deseadas. Este requisito conocido como el de la *simetría* excluye como procedimiento para equiparar puntuaciones al método de la regresión.
4. Las puntuaciones de las pruebas, una vez realizada la equiparación, deben ser totalmente intercambiables, es decir, después de la equiparación deben ser idénticas las distribuciones condicionales de las puntuaciones de cada prueba, dado un determinado nivel de rasgo o

característica que éstos evalúan. Esto implicaría que, sin importar el nivel de habilidad del sujeto, no debe tener ninguna implicación que se le aplique la prueba X o el Y. Este requisito es el más controvertido de los cuatro. En efecto, para poder satisfacer esta condición las pruebas deben ser, bien perfectamente fiables, bien estrictamente paralelos y, en cualquiera de los casos, no sería necesario equiparar.

A pesar de las fuertes restricciones que ejercen estas condiciones sobre los datos, es posible equiparar puntuaciones en muchas situaciones prácticas. Petersen, Kolen, & Hoover (1989) señalan que “probablemente no existe ningún método de equiparación que produzca puntuaciones totalmente equivalentes en dos formas de una misma prueba... (sin embargo), una equiparación aproximada de las puntuaciones en dos formas de una prueba resultará más justa que no realizar ningún tipo de equiparación, sobre todo, si las formas de la prueba difieren en dificultad” (p. 302).

Referencias

- Angoff, W. H. (1984). *Scales, norms, and equivalent scores*. Princeton, NJ: Educational Testing Service.
- Dorans, N. J. & Holland, P. W. (2000). Population invariance and the equability of test: Basic theory and the linear case. *Journal of Educational Measurement*, 37, 281-306.
- Holland, P. W. & Rubin, D. B. (1982). *Test Equating*. Academic Press.
- Kolen, M. J. (2001). Linking assessments effectively: Purpose and design. *Educational Measurement: Issues and Practice*, 20, 5-9.
- Kolen, M. J. (2004). Linking assessments: Concept and history. *Applied Psychological Measurement*, 28, 219-226.
- Herrera, A. N. (2005). *Tamaño de muestra y razón de tamaño del funcionamiento diferencial de los ítems*. Tesis doctoral no publicada, Universidad de Barcelona, Barcelona.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: LEA.
- Muñiz, J. (1997). *Introducción a la Teoría de Respuesta a los Ítems*. Madrid: Pirámide.
- Navas, M. J. (1996). Equiparación de puntuaciones. En J. Muñiz (Ed.), *Psicometría* (pp. 293-369). Madrid: Universitas, S. A.
- Petersen, N. S., Cook, L. L. & 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., & Hoover, H. D. (1989). Scaling, norming and equating. En R. L. Linn (Ed.), *Educational Measurement*, New York: Macmillan.
- Pommerich, M., Harris D. J., Hanson, B. A., & Sconing, J. A. (2004). Issues in conducting linkages between distinct tests. *Applied Psychological Measurement*, 28, 247-273.
- Skaggs, G. & Lissitz, R. W. (1986). IRT Test Equating: Relevant issues and a review of recent research. *Review of Educational Research*, 56, 495-529.
- Yin, P., Brennan, R. B., & Kolen, J. (2004). Concordance between ACT and ITED scores from different populations. *Applied Psychological Measurement*, 28, 274-289