6

Funcionamiento diferencial del ítem por sexo en alumnos de educación secundaria

Differential item functioning by sex in secondary school education students

Delia Arroyo Resino
Universidad Complutense de Madrid, España

Resumen

Muchas de las pruebas utilizadas para medir los niveles de logro no se encuentran exentas de artefactos estadísticos como es el Funcionamiento Diferencial del Ítem (DIF). Este aparece cuando estudiantes con el mismo nivel de rasgo, difieren a la hora de acertar un ítem. Desde esta perspectiva, con el fin de garantizar la calidad de los resultados, se estudió la posible existencia de DIF en los ítems y en el test de acuerdo a la variable sexo (hombres y mujeres) en una prueba de razonamiento matemático realizada por 697 estudiantes (52% hombres y 48% mujeres) de 2º de Educación Secundaria Obligatoria. La prueba estaba compuesta de 25 ítems cuya fiabilidad total fue de 0.77. Para la detección del DIF en los ítems se realizó una regresión logística y el test de verosimilitudes. Ambas pruebas indicaron la existencia de DIF pero no todos los ítems eran coincidentes. Asimismo, los criterios de Zumbo y Thomas (1997), y de Jodoign y Gierl (2001) revelaron un DIF poco perceptible. Para el estudio del DIF en el test se analizó, entre otras pruebas, la Curva Característica del Test y se calculó la ETSSD, ambos resultados mostraron que el DIF no era muy relevante a nivel del test. A la vista de estos resultados concluimos que aunque algunos ítems se encuentran en cierta medida sesgados a favor de uno de los dos sexos, el DIF es tan imperceptible que no afecta a nivel del test y por tanto las medidas de rendimiento obtenidas no se encuentran afectadas por DIF.

Palabras clave: funcionamiento diferencial, rendimiento académico, calidad de resultados.

Cita sugerida:
Arroyo Resino, D. (2018). Funcionamiento diferencial del item por sexo en alumnos de educación secundaria. En G. de la Cruz Flores (Coord.), *Experiencias educativas en el aula de Infantil, Primaria y Secundaria.* (pp. 66-76). Eindhoven, NL: Adaya Press.

Abstract
Many of the tests to measure the achievement levels are not exempted of statistical artifacts such as the Differential Item Functioning (DIF). This appears when students, with the same performance, differ when they have to guess an item. From this perspective, and with the purpose of guaranteeing the quality of the results, it was studied the existence of DIF in the items and the test in accordance with the sex variable (men and women) in a test of mathematical reasoning carried out by 697 students (52% men and 48% women) of 2nd of secondary school. The test was composed of 25 items whose total test reliability was 0.77. For the detection of the DIF in the items, a logistic regression and test of verisimilitude were made. Both tests showed the existence of DIF, but not all of the items were coincident. Moreover, the criteria of Zumbo and Thomas (1997), and Jodoign and Gierl (2001) revealed a DIF not very noticeable. For the study of the DIF in the test it was analyzed, among others, the Characteristic Curve of the Test and it was calculated the ETSSD. Both results showed that the DIF wasn't very relevant in a test level. In view of these results we conclude that, although some items are found in a certain extent biased in favor of one of the two sexes, the DIF is so unnoticeable that it doesn't affect the level of the test and thereby the obtained performance measures are not affected by DIF.

Keywords: differential functioning, academic performance, results quality.

Introducción[1]

Algunas pruebas que miden el rendimiento de los estudiantes se pueden encontrar sesgadas por la existencia de ciertos artefactos estadísticos que afectan a las propiedades psicométricas de las mismas. Entre los mismos, es posible encontrar instrumentos que presentan un funcionamiento diferencial a nivel de ítem (DIF) y/o a nivel de test (DTF). Desde el marco de la Teoría de Respuesta al Ítem (TRI) se habla de DIF cuando difieren las dos Curvas Características de los Ítems (CCIs), tras haber fijado la misma métrica en los dos grupos (Navas, 2000). Generalmente los grupos son denominados como grupo focal (F) que es el que va a ser estudiado para determinar si tiene desventaja (minoritario) y el grupo de referencia (R) que suele ser el grupo mayoritario.

Los «*Standards for Educational and Psychological Testing*», consideran que existe funcionamiento diferencial del ítem cuando, individuos con el mismo nivel de rasgo difieren, en términos promedio, en su respuesta a un ítem particular (American Educational Research Association, 2014). Acorde con esta definición, podemos afirmar que un reactivo tiene DIF cuando individuos pertenecientes a grupos distintos, pero con el mismo nivel de rasgo, tienen distintas probabilidades de responder correctamente a un

1 Desarrollo del resumen publicado en el Book of Abstacts CIVINEDU 2017 (Arroyo, 2017).

ítem, es decir, ante la existencia de DIF la probabilidad de acertar un ítem no depende de la habilidad del sujeto en el rasgo que mide el ítem. En este sentido, se pueden dar dos tipos de DIF. El DIF uniforme o unidireccional caracterizado porque en todos los niveles del rasgo siempre beneficia al mismo grupo (focal o de referencia). En este sentido, las respectivas CCIs no se cruzan en ningún nivel de la variable medida ya que una de ellas siempre muestra puntuaciones superiores, o DIF no uniforme o no unidireccional donde de acuerdo al nivel de rasgo beneficia a grupos distintos (Holland y Thayer, 1988) y en algún punto las CCIs se cruzan debido a que no siempre el mismo grupo muestra puntuaciones superiores, existe una alternancia entre los dos grupos

La existencia de DIF en instrumentos educativos supone no poder conocer el verdadero nivel de rendimiento de los estudiantes ya que el mismo se encuentra medido por una prueba con propiedades psicométricas deficientes, algo que en la mayoría de las ocasiones pasa desapercibido.

Esto puede tener consecuencias muy importantes dentro del sistema educativo, ya que muchas veces el resultado obtenido con estos instrumentos de medidas se utilizan para clasificar a los estudiantes dentro de un determinado nivel de logro.

Si esta categorización se basa en un rendimiento que se encuentra mal medido, el sistema educativo no cumplirá su función de proporcionar una educación de calidad adaptándose a las características y necesidades de los estudiantes ya que se adaptará a unos niveles de logro que son reales. En este contexto, medidas que logren identificar con mayor precisión y confiabilidad el desempeño de los estudiantes resultan ser sumamente relevantes (OCDE, 2011).

Con la finalidad de que este no curra y que las pruebas utilizadas para conocer los niveles de logro de los estudiantes, cuenten con la suficiente fiabilidad para acércanos lo máximo posible al rendimiento verdadero de los estudiantes existen diferentes procedimientos para la detección del DIF. Casi todos ellos tienen en común la consideración de la unidimensionalidad y homogeneidad de la prueba (Martínez Arias, 2005). Autores como Teresi y Jones (2013) consideran que estos procedimientos para la detección del DIF deben asumir no solo el supuesto de unidimensionalidad sino también de independencia local.

Uno de los métodos clásicos para la estimación de DIF es la regresión logística. Dicho método es parecido al método de regresión lineal, difieren en que en la regresión logística, la relación lineal se produce entre las variables predictoras y el Logit (P). El Logit (P) se estima de la siguiente manera:

$$Logit(P) = \beta_0 + \beta_1 X + \beta_2 2\, g + \beta_3 (X_g)$$

En la ecuación anterior si el ítem funciona bien y carece de DIF, $\beta_1 X$ que es la puntuación en el test, debe ser significativa. En el caso de que el parámetro β_2 resulte significativo indica la presencia de DIF uniforme y si β_3 es significativo supone la existencia de DIF no unidirrecional.

En definitiva, en la regresión logística se estudia si un ítem posee DIF comparando un modelo donde solo tiene como predictor la variable X, con otro que tiene como predictores la variable g y X_g a partir del estudio del estadístico X_2. La hipótesis nula que se pone a prueba es que no hay DIF (ya que el modelo 1 no es significativamente peor que el modelo completo).

Generalmente cuando se utiliza la regresión logística, y existe DIF, para conocer el tamaño del efecto del mismo se analizan los cambios que se producen en el r^2 de Nagelkerke. La interpretación de dichos valores se realiza mediante los criterios de Zumbo y Thomas (ZT) y Jodoign y Gierl (JG). Según ZT cuando el cambio en r^2 oscila entre 0 y .13 se considera que el DIF no es perceptible, cuando r^2 tiene un valor comprendido entre .13 y .26 el DIF es moderado y si el valor es mayor a .26 se considera que el DIF es bastante perceptible. Los criterios de JG son algo más rigurosos ya que se considera que el DIF es no perceptible cuando r^2 oscila entre 0 y .035. Entre .035 y .07 el DIF es moderado y cuando el valor de r^2 es mayor a .07 se habla de un DIF muy perceptible.

Desde la TRI una de las técnicas más utiliza para la detección de DIF es el Test de Razón de Verosimilitudes. Un test propuesto por Thissen (2001) donde se pone a prueba si los parámetros de un ítem son o no son iguales en los dos grupos. Para ello se basa en la razón de verosimilitudes:

$$LR=\left[\frac{L_1}{L_2}\right]$$

En la ecuación anterior, L_1 (modelo parsimonioso) se refiere a los parámetros del ítem contrastados iguales a través de los grupos y L_2 (modelo complejo) se refiere a los parámetros del ítem contrastados diferentes a través de los grupos. Estos parámetros serán más diferentes cuando el DIF sea más perceptible.

Para poder compararlos ambos modelos L_1 y L_2 es necesario que se encuentren anidados. Existen distintas estrategias de comparación de modelos, según lo que asumamos sobre el resto de los ítems (distintos al ítem del que se analiza el DIF). Por ejemplo, está la estrategia donde todos los demás ítems son invariantes, ninguno de los demás ítems es invariante, la purificación, entre otras (Kim y Kolen, 2007).

El problema de utilizar dichas técnicas para la realización de comparaciones es que al hacer tantas comparaciones como ítems, por puro azar al menos un 5% de las comparaciones saldrían significativas (Kim y Kolen, 2007). Por ello es común el uso de métodos de corrección por comparaciones múltiples como es el método de Benjamini-Hochberg (1995), que permite la obtención de un valor de p ajustado. Dicha corrección se basa en ordenar el valor las p de menor a mayor (rango 1 a n) y posteriormente multiplicar cada p por n/rango.

Además de la existencia de un funcionamiento diferencial a nivel de ítems, este también se puede dar a nivel de test (DTF), en este sentido, para la detección del mismo inicialmente se utilizan pruebas como la t-student para ver si existe diferencias entre ambos grupos. Pero también es común el estudio de la Curva Característica del Test (CCT) con el fin de comparar los niveles de rendimiento en ambos grupos cuando tienen el mismo nivel de logro.

Otro indicador utilizado para la detección de DTF es el *Signed Test Score Difference (STSD),* el STDS como se muestra en la ecuación siguiente, hace referencia a la diferencia en promedio de la puntuación esperada en el test para los sujetos del grupo focal.

$$STSD = \sum_{q} g\left(\theta_q \middle| \mu_{2(f)}, \sigma_{2(f)}\right)\left[CCT_{1(F)}\left(\theta_q\right)\right]$$

Si tiene un valor próximo a 0 es que o bien no hay DIF, o bien el DIF es no unidireccional y se cancela a través de las thetas a nivel de ítem, o el DIF es unidireccional pero también se cancela a través de los ítems a nivel del test.

Finalmente, otra prueba vinculada a la anterior es la *Expected Test Score Standardized Difference in the Sample (ETSSD)* que alude a la diferencia promedio en la puntuación esperada en el test estandarizado para los sujetos del grupo focal.

$$ETSSD = \frac{STSD}{SD}$$

Este indicador se interpreta mediante una métrica en unidades de desviación típica, por lo que para su interpretación se aplica los criterios de Cohen (1988) donde se considera la existencia de un DTF pequeño con valores de .2, medio cuando tiene un valor entorno a .5 y grande con valores de .8 o superiores. Si además el valor de dicho coeficiente es positivo el DTF se considera que se da a favor del grupo focal.

Metodología

Diseño

De acuerdo con Kerlinger y Lee (2002) dicho trabajo se englobaría dentro de los diseños no experimentales, ya que no se manipula ninguna variable indpendiente dado que tan sólo se investiga la posible existencia de diferentes artefactos estadísticos en una prueba de rendimiento.

Objetivo

Estudiar si existe un funcionamiento diferencial a nivel de ítems (DIF) y del test (DTF) de acuerdo a la variable sexo (hombres y mujeres).

Participantes

La muestra fue incidental y estuvo compuesta por 697 estudiantes (52% hombres y 48% mujeres), con un error muestra de .08, de 2º de Educación Secundaria Obligatoria (ESO) de distintos colegios de la Comunidad de Madrid.

Instrumento

Se trata de una prueba de razonamiento matemático formada por 25 ítems de respuesta múltiple (4 opciones) que se dicotomizaron en errores (0) y aciertos (1). Con esta prueba,

utilizada en una evaluación ordinaria, por distintos docentes de Educación Secundaria, se pretendía evaluar fundamentalmente el razonamiento del estudiante ante problemas matemáticos, así como su aplicación práctica a la vida cotidiana. Por tanto, su objetivo no es tanto verificar directamente el conocimiento que tienen los estudiantes sobre contenidos, hechos o datos concretos, sino más bien analizar su nivel para examinar o valorar información relacionada con la resolución de problemas contextualizados en situaciones que le pueden resultar más o menos cotidianas. Estas pruebas buscan que el estudiante ponga en marcha diferentes procesos referidos a diferentes contenidos en diversos contextos como indicador de su nivel competencial.

La fiabilidad total del test fue de 0,77. Algunos ejemplos de estos ítems son el ítem 2. Si acabas de leer las 2/5 de un texto de 55 líneas, ¿Cuántas líneas te quedan por leer?, a)33, b)22, c)11 y d)30. El Ítem 5: La solución a la siguiente operación (-10) / (-10) es: a)-20, b) +100, c) +1 y d)-1.

Procedimiento

Previo a la detección del funcionamiento diferencial del ítem se comprobó el supuesto de unidimensionalidad y de independencia local del test. Para comprobar la unidimensionalidad se aplicó un análisis paralelo y un análisis factorial exploratorio. Para la independencia local se utilizó el estadístico *chi-cuadrado* estandarizado. Posteriormente con la finalidad de estudiar si existían diferencias significativas entre las medias del grupo de hombres y de mujeres se realizó una prueba *t-student* donde el grupo de referencia era los hombres, por tener un mayor tamaño muestral, y el focal eran las mujeres. Una vez comprobada la diferencia de medias se procedió a la detección del funcionamiento diferencial de los ítems mediante una regresión logística interpretada con los criterios de Zumbo y Thomas (ZT) y Jodoign y Gierl (JG). Asimismo, también se realizó el test de verosimilitudes con la estrategia de fijar a todos los ítems como invariantes y además se aplicó la corrección de Benjamini-Hochberg (1995) con el fin de evitar el problema de las comparaciones múltiples. Finalmente, para comprobar si existía un funcionamiento diferencial a nivel del test se analizó las Curvas Características del Test de los dos grupos, y se calculó la diferencia con signo de la puntuación del test (STDS) y la diferencia promedio de la diferencia esperada (ETSSD).

Análisis de datos

Los análisis de datos se realizaron con el software estadístico R-Studio. El nivel de significación utilizado fue .05.

Resultados

Unidimensionalidad e independencia local

Para la comprobación del supuesto de unidimensionalidad se realizó un análisis paralelo. Como podemos observar en la siguiente tabla se retuvieron dos factores.

Tabla 1. Retención de factores

Items	Autoevaluadores observadores	Autoevaluadores aleatorios
1	8.006	1.57
2	1.792	1.501
3	1.415	1.445
4	1.383	1.392
5	1.242	1.348
6	1.188	1.318
7	1.179	1.293
8	1.139	1.262
9	1.090	1.231
10	1.065	1.199

Además del análisis paralelo también se realizó un análisis factorial confirmatorio para ver si se podía mantener la estructura unifactorial. De esta manera se formuló un modelo donde los 25 ítems cargaban en un único factor. Con esta estructura factorial se obtuvo un valor de CFI de .952 y de TLI de .950.

La independencia local entre los ítems se contrastó mediante el estadístico chi-cuadrado estandarizado, el cual reveló la existencia de dependencia local únicamente entre los ítems 2 y 4 ya que el valor de chi-cuadrado estandarizado fue superior a 10.

Análisis de detección del Funcionamiento Diferencial del Ítems (DIF)

Para ver si existía DIF se realizó una prueba T. Como p<0,05 se rechazó la hipótesis de igualdad de medias, mostrando en el grupo de hombres una media superior a la de las mujeres (mujeres = 16,96 y hombres =18,81).

Una vez realizado la prueba T se realizó una regresión logística la cual reveló la posible existencia de DIF en los ítems 4, 5, 6, 8, 9, 10, 11, 18, 22, 24 y 25.

A modo de ejemplo se muestran las Curvas Características del Ítem 4 y 6. Como observamos en la figura 1, el ítem 4 muestra DIF a favor del grupo de referencia (hombres) en los niveles bajos y medios de logro. Según se va incrementado el nivel de rasgo las probabilidades de acertar el ítem se van igualando en ambos grupos. En el caso del ítem 6, con niveles de rasgo bajo el grupo focal (mujeres) tienen mayor probabilidad de acertar el ítem y en los niveles altos se iguala.

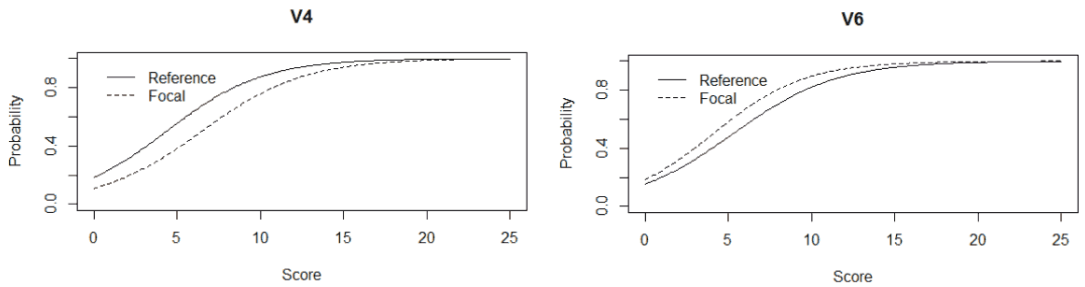

Figura 1. Ítems con DIF

Para estudiar la magnitud de DIF en los diferentes ítems detectados en la regresión logística se utilizaron los criterios de Zumbo y Thomas (ZT) y Jodoign y Gierl (JG). De acuerdo al criterio de ZT todos los ítems que podían tener DIF, presentaban un DIF no perceptible. Sin embargo, siguiendo el criterio JG el único ítem que tenía un DIF no perceptible era el 18. Los ítems 4, 24 y 25 se consideraban que tenían un DIF alto y el resto de reactivos (5, 6, 8, 9, 10, 11) contaban con unos valores de DIF moderados.

Otra prueba realizada para el estudio del DIF en los ítems fue el test de verosimilitudes utilizando la estrategia de dejar a todos los ítems invariantes en los parámetros a excepción de aquellos que presentaban DIF. Con dicha prueba los ítems que mostraron DIF fueron el: 4, 5, 10, 18 y 24 a favor del grupo de referencia y el 9, 11, y 22 a favor del grupo focal.

Finalmente, para comprobar la significatividad del DIF y evitar los problemas de comparabilidad se calculó la p-ajustada de acuerdo a la corrección de Benjamini-Hochberg (1995). La tabla 5 muestra la existencia DIF en los mismos ítems que se había detectado anteriormente con la estrategia invariante.

Tabla 2. Utilización de la p-ajustada para la detección del DIF

Ítems	X2	df	p	p ajustada
1	5.49	2	0.064	0.134
2	0.14	2	0.930	0.961
3	0.77	2	0.679	0.772
4	17.96	2	**0.000**	**0.001**
5	15.18	2	**0.000**	**0.003**
6	7.28	2	**0.026**	0.073
7	0.08	2	0.961	0.961
8	5.69	2	0.058	0.132
9	10.98	2	**0.004**	**0.018**
10	9.69	2	**0.008**	**0.025**
11	10.94	2	**0.004**	**0.018**
12	3.70	2	0.157	0.302
13	0.36	2	0.834	0.907
14	1.40	2	0.497	0.772
15	0.79	2	0.674	0.772
16	1.85	2	0.397	0.662
17	0.87	2	0.648	0.772
18	10.37	2	**0.006**	**0.020**
19	0.83	2	0.662	0.772
20	3.05	2	0.218	0.389
21	1.21	2	0.547	0.772
22	24.83	2	**0.000**	**0.000**
23	0.96	2	0.620	0.772
24	19.92	2	**0.000**	**0.000**
25	6.91	2	**0.032**	0.079

Análisis de detección del Funcionamiento Diferencial del Test (DTF)

Para estudiar si había un funcionamiento diferencial a nivel del test, en primer lugar, se realizó un estudio de las Curvas Características del Test. Como podemos observar en la figura 2, aunque la diferencia entre ambos grupos no es muy significativa, se puede apreciar que el test en su conjunto es más fácil para los hombres (grupo 1) que para las mujeres (grupo 2) ya que, sobre todo en los niveles bajos y medios del rasgo, los hombres cuentan con una mayor probabilidad de aciertos, que se va igualando con la probabilidad de acierto de las mujeres en los niveles altos de logro.

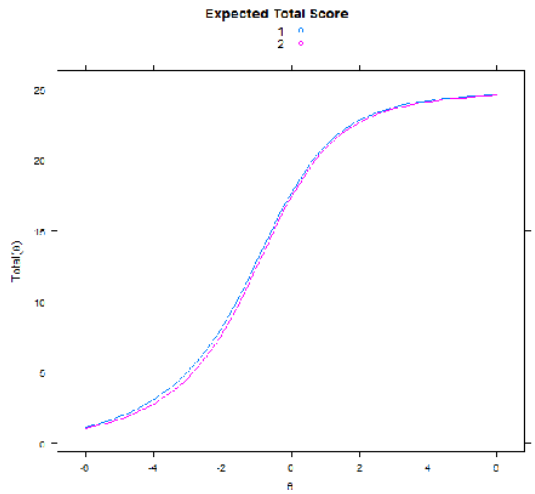

Figura 2. Curva Característica del Test

En cuanto a la diferencia en promedio de la puntuación esperada en el test para los sujetos del grupo focal (STDS) se obtuvo un valor de .03. Al ser dicho valor tan cercano a 0 se consideró que el DIF no era perceptible a nivel del test. El valor del estadístico ETSSD también fue muy bajo (.07) por lo que siguiendo los criterios de Cohen se reafirma la inexistencia de DIF en el test.

Conclusiones

El objetivo de dicho trabajo era estudiar la existencia de DIF a nivel del ítem y del test de acuerdo a la variable sexo, ya que la existencia de dicho artefacto estadístico en los instrumentos de medida puede conllevar a estimaciones de los niveles de logro de los estudiantes imprecisos.

Previo al cumplimiento de dicho objetivo se comprobó la unidimesionalidad y la independencia local. Respecto a la unidimensionalidad aunque se retuvieron dos factores, los datos sí que se ajustaron al modelo unifactorial con unos índices adecuados, ya que según Hu y Bentler (1999) valores de CFI y TLI superiores a .95 indican un buen ajuste del modelo a los datos. Además, el valor del coeficiente RMSEA fue de .02, al ser dicho

valor inferior a .5 podemos hablar de un buen ajuste (Browne y Cudeck, 1993). Por lo tanto, aunque dicha prueba no contó con una unidimensionalidad en sentido estricto, sí que los datos ajustaron al modelo unifactorial con unos valores de índices aceptables. Del mismo modo, también se mantuvo el supuesto de independencia local ya que solo se encontró dependencia local entre el ítem 2 y 4, esta dependencia puede ser debido a que los contenidos de los dos reactivos son muy similares ya que ambos hacen referencia a operaciones con fracciones.

En cuanto a la detección de DIF a nivel de ítem, la regresión logística reveló la posibilidad de existencia de DIF en los ítems 4, 5, 6, 8, 9, 10, 11, 18, 22, 24 y 25, aunque no se trataba de un DIF muy perceptible en todos ellos de acuerdo a los criterios de ZT y JG. En el caso del test de la razón de verosimilitudes mostró también la posible existencia de DIF en los mismos ítems anteriores, a excepción de los ítems 6, 8 y 24. Por lo que no hubo un acuerdo unánime entre ambas pruebas.

En cuanto al estudio del funcionamiento diferencial a nivel del test las distintas pruebas realizadas mostraron la inexistencia del mismo. Por lo tanto, a la vista de estos resultados se concluye que aunque existe DIF en ciertos ítems, tomando el test en su conjunto, el mismo no se encuentra afectado por un funcionamiento diferencial; por ende, los resultados obtenidos con dicha prueba no se encuentran sesgados a favor de ningún grupo.

Referencias

American Educational Research Association, Américan Psicological Association & National Council on Measurement in Education (2014). *Standards for educational and psychological testing.* Washington: American Educational Research Association, Américan Psicological Association & National Council on Measurement in Education.

Benjamini, Y., y Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, 57*(1), 289-300.

Browne, M.W., y Cudeck, R. (1993). Alternative ways of assessing model fit. En K.A. Bollen y J.S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Newbury Park, CA: Sage.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Second Edition. Hillsdate, NJ: LEA.

Hu, L., y Bentler, P. (1999). Cutoff criteria for fit indices in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1-55.

Holland, P. W., y Thayer, D. T. (1988). Differential Item Performance and the Mantel Haenszel Procedure. En H. Wainer y H. I. Braun (Eds.), *Test Validity* (pp. 129-145). Hillsdale, New Jersey: Lawewnce Erlbaum Associates.

Jodoin, M. G., y Gierl, M.J. (2001). Evaluating Type I error and power rates using an effect size measure with logistic regression procedure for DIF detection. A*pplied Measurement in Education, 14*, 329–349.

Kerlinger, F. N., y Lee, H. B. (2002). *Investigación del comportamiento. Métodos de investigación en ciencias sociales* (4ª ed.). Mexico: McGraw-Hill.

Kim, S., y Kolen, M. J. (2007). Effects on Scale Linking of Different Definitions of Criterion Functions for the Irt Characteristic Curve Methods. *Journal of Educational and Behavioral Statistics, 32*(4), 371-397.

Martínez-Arias, R. (2005). *Psicometría: teoría de los tests psicológicos y educativos.* Madrid: Síntesis.

Navas, M. J. (2000). Equiparación de puntuaciones: exigencias actuales y retos de cara al futuro. *Metodología de las Ciencias del Comportamiento*, *2*(2), 151-165.

OCDE (2011). *La medición del aprendizaje de los alumnos: Mejores prácticas para evaluar el valor agregado de las escuelas*, OECD Publishing. http://dx.doi.org/10.1787/9789264090170-es.

Teresi, J. A., y Jones, R. N. (2013). Bias in psychological assessment and other measures. En K. F. Geisinger (Ed.), *APA Handbok of Testing and Assessment in Psychology* (pp. 139-164). Washington D.C: American Psychological Asociation.

Thissen, D. (2001). IRTLRDIF v.2.0b: *Software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning* [Computer software and manual]. University of North Carolina at Chapel Hill.

Thissem, D., Steinberg, L., y Kuang, D. (2002). Quick and easy implementation of the Benjamini-Hochberge procedure for controlling the false positive rate in multiple comparisons. *Journal of Educational and Behavioral Statistics, 27*(1), 27-77.

Zumbo, B. D., y Thomas, D. R. (1997) *A measure of effect size for a model-basedapproach for studying DIF*. Working Paper of the Edgeworth Laboratory for Quantitative Behavioral Science. University of Northern British Columbia: PrinceGeorge, B.C

Delia Arroyo Resino, licenciada en Pedagogía por la Universidad Complutense de Madrid con premio extraordinario de fin de carrera. Ha realizado el Máster de Estudios Avanzados de Pedagogía y el Máster Interuniversitario de Metodología de las Ciencias del Comportamiento y la Salud. Actualmente se encuentra realizando el doctorado en Educación, concretamente en el departamento de Métodos de Investigación y Diagnóstico en Educación (MIDE) con una beca competitiva otorgada por el Ministerio de Educación de la Comunidad de Madrid para la Formación del Profesorado Universitario (FPU).

ISBN 978-94-92805-06-5