SUMMA PSICOLÓGICA UST
2013, Vol. 10, Nº 2, 49-56

ISSN 0718-0446 (impresa)

Criterios para la eliminación de ítems de un Test de Analogías Figurales

Criteria for eliminating items of a Test of Figural Analogies

Diego Blum[1] Sofía Auné
María Silvia Galibert Horacio Attorresi
Instituto de Investigaciones de la Facultad de Psicología de la Universidad de Buenos Aires.

(Rec: junio 2013 – Acep: octubre 2013)

Resumen

En el presente artículo se describen los pasos que permiten considerar la eliminación de dos de los reactivos de un Test de Analogías Figurales (TAF). Se explican las bases generales de su análisis psicométrico tanto desde la Teoría Clásica de Tests (TCT) como la Teoría de Respuesta al Ítem (TRI). Se detalla el proceso de eliminación de reactivos basado en la obtención de los índices de dificultad y discriminación de los ítems desde la TCT, así como el análisis de la unidimensionalidad. También se toman criterios basados en la estimación de los parámetros *a*, *b* y *c* de cada ítem según el Modelo Logístico de Tres Parámetros de la TRI, así como la determinación del ajuste de cada reactivo sobre la base de dicho modelo logístico. Se detallan las características desfavorables de algunos ítems del TAF y se discuten las decisiones que llevan al mencionado proceso de eliminación.

Palabras clave: eliminación de ítems, Test de Analogías Figurales, Teorías de los Tests

Abstract

This paper describes the steps taken to eliminate two of the items in a Test of Figural Analogies (TFA). The main guidelines of psychometric analysis concerning Classical Test Theory (CTT) and Item Response Theory (IRT) are explained. The item elimination process was based on both the study of the CTT difficulty and discrimination index, and the unidimensionality analysis. The *a*, *b*, and *c* parameters of the Three Parameter Logistic Model of IRT were also considered for this purpose, as well as the assessment of each item fitting this model. The unfavourable characteristics of a group of TFA items are detailed, and decisions leading to their possible elimination are discussed.

Key words: item elimination, Test of Figural Analogies, Test Theories

[1] Correspondencia a: Diego Blum. Dirección Postal: Anchorena 1.169 3ºB (1425), Ciudad de Buenos Aires. Teléfono: (005411) 5778-1813. E-mail: blumworx@gmail.com.

Introducción

La construcción de pruebas psicológicas es un proceso complejo que consta de varias etapas, las cuales involucran el entendimiento conceptual y operacional del constructo a medir (Martínez Arias, Hernández Lloreda, & Hernández Lloreda, 2006; Prieto & Delgado, 1996). A través del establecimiento del marco teórico, se sabe qué conductas observables son el mejor referente de la expresión del constructo y qué tareas específicas de la prueba pueden llegar a suscitarlas. Este conocimiento es clave para la construcción de los reactivos de la prueba, los cuales se convierten idealmente en indicadores conductuales del constructo. De esta forma, se definen el número de ítems, el tipo de ítems (verbales, gráficos, pictóricos, etc.), el tipo de respuesta (cerrada o de elección múltiple, de ordenamiento de material, abierta, etc.), el orden de los reactivos y la delimitación de la consigna. Asimismo, para una prueba de habilidad, suelen utilizarse ítems con respuestas apropiadas e inapropiadas (por ejemplo la respuesta dicotómica: *acierto* o *no-acierto*).

Desde un enfoque psicométrico, la medición del Razonamiento Analógico (RA) requiere construir pruebas de habilidad cuyos ítems sean indicadores fieles del constructo y lo midan con precisión. Existe un sinnúmero de pruebas que miden, ya sea directa o indirectamente, la capacidad de razonar analógicamente con relaciones visuoespaciales, tales como el Test de Matrices Progresivas de Raven (Raven, Court, & Raven, 1993) y el Test de Inteligencia No-Verbal versión 2 (Test of Non-Verbal Intelligence 2, TONI 2) de Brown, Sherbenou & Johnsen (2000). Incluso existen trabajos en Argentina (Blum, Abal, Galibert, & Attorresi, 2011; Blum, Galibert, Abal, Lozzia, & Attorresi, 2011).

Luego de su construcción, el test se pone a prueba definitivamente frente a los estudios empíricos. Se administra a sucesivas muestras y se recaba información para determinar las características psicométricas del test y de sus ítems. En esta etapa es cuando se toma conocimiento real acerca del funcionamiento de los reactivos, teniendo en cuenta que muchos de ellos posiblemente no funcionen de manera adecuada o acorde con lo esperado. Por esta razón es que suele sugerirse la administración de una cantidad mucho mayor de reactivos en función de la cantidad que se espera conservar para el test definitivo. En otros términos, muchos de los ítems previamente construidos deben eliminarse o modificarse luego de que, frente a sucesivas muestras, prueban ser ineficaces para medir el constructo.

Los análisis más frecuentes que se realizan para tal objetivo provienen de la Teoría Clásica de Tests (TCT), aunque recientemente se ha observado un gran avance de aquellos modelos conocidos con el nombre genérico de Teoría de Respuesta al Ítem (TRI). Se considera que la TRI es un excelente complemento del enfoque clásico, ya que en determinadas circunstancias ofrece soluciones a problemas mal resueltos en el mismo (Muñiz, 2010). Se han desarrollado investigaciones con tests de razonamiento no verbal que incluyeron los aportes conjuntos de la TCT y la TRI, tales como las mencionadas por Susan Embretson acerca de su Test de Razonamiento Abstracto (Abstract Reasoning Test, ART. Embretson & Reise, 2000) y los estudios realizados por Raven, Raven & Court (1991) entre los que se cuenta la interpretación visual de las Curvas Características de los Ítems.

La TCT en particular, descompone el puntaje observado de un individuo a un test, en la suma de una puntuación verdadera (nivel real que la persona posee del rasgo evaluado) y de un error de medida (variable aleatoria). Una de las medidas más conocidas de la TCT es el Coeficiente de Confiabilidad, para cuya estimación pueden analizarse los datos de una sola muestra sin requerir las medidas de un eventual retest o de un test paralelo. A esto se lo conoce como el análisis de la congruencia interna. El método más utilizado con este fin es el Coeficiente α de Cronbach (1951). Los coeficientes de congruencia interna están basados en la búsqueda de intercorrelación entre los ítems de un test. Esto quiere decir que cualquier ítem cuya correlación con el puntaje total de la prueba (corregido sin considerar el ítem) sea baja, nula o negativa generará una reducción de la confiabilidad. En otras palabras, tales reactivos provocan la reducción del valor de α y esto suele traer por consecuencia su eliminación, con el propósito de que α sea más elevado entre aquellos reactivos no eliminados. Los estudios empíricos de Borg (1963) para tamaños de muestra iguales o mayores que 100, mostraron que en general las correlaciones iguales o mayores que .35 son estadísticamente significativas al 1% (Cohen & Manion, 2002).

Por otro lado, un criterio muy empleado para estudiar la validez del test desde el marco de la TCT es el análisis de su dimensionalidad, es decir, el conocimiento del número de factores que determinan las respuestas a los ítems, en tanto dichos factores explican mejor la conducta a medir. Uno de los métodos más utilizados con este fin es el análisis factorial, gracias al cual se extraen aquellos grupos de ítems más correlacionados entre sí debido a que remiten a factores no-observables

(Argibay, 2006). Para ítems de respuesta dicotómica suele recomendarse un análisis factorial con la matriz de correlaciones tetracóricas (García-Cueto & Fidalgo, 2005). Mediante análisis sucesivos de la dimensionalidad, pueden identificarse reactivos cuya eliminación permite arribar a estructuras factoriales más convenientes en función de lo que se quiere medir. Si el fin es que la prueba mida un solo atributo, la eliminación de determinados ítems se realiza en favor del aumento de la unidimensionalidad, al permitir que los ítems posean una mayor saturación factorial en el primer autovalor.

A lo largo de la historia de la Psicometría, la Teoría de Respuesta al Ítem (TRI) ha sobresalido por la magnitud de sus aportaciones a la calidad de la medición psicológica (Allen & Yen, 1979; Gulliksen, 1950; Hambleton, Swaminathan, & Rogers, 1991; Lord, 1952, 1980; McDonald, 1999; Muñiz, 1997; Santisteban, 1990). Estos estudios parten de la premisa de que la respuesta de un individuo ante un ítem puede predecirse y explicarse a partir de una o múltiples variables inobservables denominadas rasgos latentes. Una de las ventajas que posee la TRI por sobre la TCT es que permite expresar las propiedades del test en función de la aditividad de las propiedades de los ítems que lo componen, es decir, que se independiza de los instrumentos particulares de medición (tests) y toma a cada ítem como una unidad de análisis básica. Los autores que formularon los modelos de la primera generación de la TRI (Birnbaum, 1968; Rasch, 1960), utilizaron los Modelos de Uno, Dos y Tres Parámetros (ML1P, ML2P y ML3P respectivamente) para la descripción formal de los ítems. Dichos parámetros se refieren al nivel de dificultad del ítem (*b*), su potencia discriminatoria (*a*) y la posibilidad de que un individuo de bajo nivel de habilidad acierte el ítem, también conocido como parámetro de pseudoazar (*c*). Por otro lado, la habilidad (θ) es la estimación del nivel del rasgo latente propio de cada individuo, que es el equivalente a la puntuación verdadera de la TCT.

Se evidencia el ajuste de un ítem al modelo cuando las proporciones de aciertos de los sujetos para cada nivel del rasgo del ítem tienden a parecerse a las proporciones teóricas esperadas (Muñiz, 1996). En otros términos, si la Curva Característica del Ítem (CCI) empírica y la teórica o esperada tienden a coincidir, entonces es esperable encontrar ajuste. Para formalizar el estudio del ajuste, suele recurrirse a la prueba χ^2 de calidad de ajuste (Embretson & Reise, 2000). El χ^2 es el estadístico de contraste que pone a prueba la hipótesis nula de que la diferencia entre los datos empíricos y los esperados es cero. El psicómetra tiene como opción eliminar aquellos ítems que no ajustan a un determinado modelo logístico, o simplemente modificar su estructura interna para continuar estudiando su ajuste posteriormente. También es posible adoptar como criterio la eliminación de reactivos con niveles de *a* reducidos, los cuales son poco deseables; éste es un razonamiento similar al que se establece con las correlaciones ítem-total de la TCT. Un último criterio de eliminación desde la TRI puede basarse en el análisis de los parámetros de pseudoazar, ya que es indeseable encontrar un ítem cuyo *c* posea un valor más alto que lo esperable teniendo en cuenta la cantidad de opciones de respuesta.

Por lo tanto, los criterios siguientes podrían ser más que suficientes para fundar la decisión de eliminar un ítem de un test de habilidad con respuesta dicotómica: correlación de este ítem con el puntaje total corregido menor a .35, aumento del α de Cronbach cuando se lo calcula sin el reactivo, la mejoría de la unidimensionalidad cuando se quita el ítem, su falta de ajuste a un modelo logístico unidimensional de la TRI, un parámetro *a* reducido y un parámetro *c* elevado. Pero para asegurar que la eliminación del ítem esté correctamente fundada, muchos de estos criterios desfavorables deberían presentarse juntos en un mismo reactivo. El establecimiento de un punto de corte, en el sentido de qué cantidad de atributos desfavorables contar para considerar tal eliminación y a cuáles brindar mayor importancia, es desarrollado en el presente artículo. Para ello, se hará referencia a la calibración de los ítems de un Test de Analogías Figurales (TAF) en sus diversas versiones y bajo ambas teorías de los tests. Este estudio representa una continuación respecto de las investigaciones de Blum, Abal et al. (2011) y de Blum, Galibert et al. (2011).

Método

Se tomaron dos muestras depuradas para esta investigación. La primera de ellas consideró los protocolos de 475 estudiantes de la Facultad de Psicología de la Universidad de Buenos Aires, mientras que la segunda comprendió 1129 protocolos de alumnos de la Facultad de Arquitectura, Diseño y Urbanismo de la misma Universidad. A los individuos se les administró un TAF con 36 reactivos de seis opciones de respuesta cada uno, con una sola correcta, sin un tiempo límite estricto de concreción de la prueba. Para acceder a los antecedentes de este instrumento, una descripción en extenso del mismo puede encontrarse en Blum, Abal

et al. (2011) y Blum, Galibert et al. (2011). La primera versión del TAF se tomó a los alumnos de Psicología, cuyos resultados se transcribieron a la matriz de datos conocida como M1. Luego se confeccionó una segunda versión de este test modificando 15 ítems respecto de la versión anterior y se la administró a la segunda muestra. Sus resultados se transcribieron a la matriz de datos conocida como M2. Por último, se unificaron los datos correspondientes a los 21 ítems comunes a ambas matrices para configurar la matriz conocida como M3, la cual contiene un total de 1604 casos.

Algunos reactivos de la segunda versión del TAF cambiaron su orden de presentación respecto de la primera versión. La Tabla 1 incluye las correspondencias entre los reactivos de ambas muestras. La columna 'TAF piloto' muestra la numeración de los reactivos según la versión administrada en Psicología. La columna "¿Modificó?" señala si el ítem de la primera columna fue o no modificado en su contenido antes de administrarlo a la segunda muestra. La columna "TAF revisado" revela la nueva ubicación del ítem según la versión administrada a la segunda muestra

Tabla 1

Correspondencias entre los reactivos de ambas muestras

TAF piloto	¿Modificó?	TAF revisado	TAF piloto	¿Modificó?	TAF revisado
1	Sí	1	19	No	31
2	No	2	20	Sí	20
3	Sí	3	21	No	21
4	No	4	22	No	22
5	No	5	23	No	23
6	Sí	30	24	Sí	18
7	No	7	25	No	25
8	No	8	26	No	26
9	Sí	9	27	No	27
10	No	10	28	Sí	28
11	Sí	11	29	No	29
12	Sí	36	30	Sí	12
13	Sí	19	31	Sí	13
14	No	14	32	No	32
15	No	15	33	No	33
16	No	16	34	No	34
17	Sí	17	35	No	35
18	Sí	24	36	Sí	6

Para el análisis de los datos se utilizó, entre otros programas, BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996). Con el mismo se estudió el ajuste global y de cada ítem al ML3P por medio de la prueba χ^2 de calidad de ajuste. También se obtuvieron los índices de la TCT y se estimaron los parámetros de la TRI descritos en la introducción.

Resultados

Tal como desarrollaron Blum, Abal et al. (2011) y Blum, Galibert et al. (2011), el TAF posee una buena calidad psicométrica tanto desde la TCT como desde la TRI. A continuación se detallan las características de aquellos ítems cuya calidad psicométrica es deficiente en función de los estudios realizados.

1. El análisis de componentes principales con la matriz de correlaciones tetracóricas permite corroborar la unidimensionalidad del test en las tres matrices de datos, ya que es evidente la existencia de un factor dominante que determina las respuestas. En cuanto al estudio de las saturaciones factoriales en M2, se verifica una mejoría de la unidimensionalidad al quitar los ítems 5 y 22 del análisis, puesto que dichas saturaciones pasan a ser altas sólo en el primer autovalor (mayores a .40).
2. Todos los α de Cronbach resultan elevados y la eliminación de cada ítem afecta directamente a dicho coeficiente reduciendo su valor. La excepción la tiene el ítem 5 de M2, puesto que al eliminarlo, α se incrementa en 13 cienmilésimos, es decir, el efecto sobre α es prácticamente nulo.
3. Una gran cantidad de correlaciones ítem-total corregido puntúa por encima de .35 en las tres matrices de datos, excepto para el ítem 20 de M1, el ítem 5 de M2 y los ítems 5 y 22 analizados en M3. Vale aclarar que el orden de presentación que ocupan estos dos últimos ítems en la primera versión del TAF, es el mismo que el que ocupan en la segunda versión; por lo tanto, aquí se los menciona en ese mismo orden al considerar M3.
4. Los parámetros de discriminación estudiados desde la TRI son generalmente elevados. Los autores de este trabajo consideran como punto de corte un $a < .65$ para determinar parámetros de discriminación reducidos. De acuerdo con este corte, aquellos reactivos con a reducido son el ítem 5 y el 20 de M1, el ítem 5 de M2 y el ítem 5 de M3.
5. A nivel global, los parámetros de pseudoazar poseen niveles levemente inferiores a .17, lo cual es esperable considerando que los ítems poseen seis opciones de respuesta. Los autores del presente artículo consideran un $c \geq .20$ como poco deseable. Aquellos reactivos con un $c \geq .20$ son los ítems 17, 18, 24 y 28 de M1, los ítems 12, 13, 17, 26 y 29 de M2 y los ítems 21, 26 y 32 analizados en M3. Con respecto a los ítems 17 y 26, los cuales poseen el mismo orden de presentación en las dos versiones del TAF, puede comprobarse que ambos poseen un $c \geq .20$ en más de una matriz de datos.
6. Por último, gran parte de los reactivos ajusta en forma individual al ML3P considerando un nivel del 5% en las tres matrices de datos. Aquellos reactivos que no ajustan al ML3P son los ítems 5, 11 y 18 de M1 y los ítems 6, 11, 15 y 17 de M2. Esto quiere decir que el único ítem que no ajusta al ML3P en más de una matriz de datos es el 11, el cual no varía en su orden de presentación a través de las versiones del TAF.

Todas las características descritas se resumen en la Tabla 2:

Tabla 2
Criterios para la eliminación de ítems

		M1			M2			M3		
Ítem en orden TAF piloto	Ítem en orden TAF revisado	Si se elimina, no reduce α	No ajusta al ML3P	Correlación Ítem-Test < .35	Si se elimina, no reduce α	Si se elimina, mejor unidim.	No ajusta al ML3P	Correlación Ítem-Test < .35	Si se elimina, no reduce α	No ajusta al ML3P	Correlación Ítem-Test < .35
			$a < 0.65$	$c \geq .20$			$a < 0.65$	$c \geq .20$			$a < 0.65$	$c \geq .20$
1	1											
2	2											
3	3											
4	4											
5	5	x		x	x	x		x	x		x	x
6	30											
7	7											
8	8											
9	9											
10	10											
11	11	x				x						
12	36											
13	19											
14	14											
15	15					x						
16	16											
17	17			x		x		x				
18	24	x		x								

		M1			M2			M3		
Ítem en orden TAF piloto	Ítem en orden TAF revisado	Si se elimina, no reduce α	No ajusta al ML3P	Correlación Ítem-Test < .35	Si se elimina, no reduce α	Si se elimina, mejor unidim.	No ajusta al ML3P	Correlación Ítem-Test < .35	Si se elimina, no reduce α	No ajusta al ML3P	Correlación Ítem-Test < .35
			$a < 0.65$	$c \geq .20$			$a < 0.65$	$c \geq .20$			$a < 0.65$	$c \geq .20$
19	31											
20	20		x	x								
21	21											x
22	22					x					x	
23	23											
24	18			x								
25	25											
26	26						x					x
27	27											
28	28			x								
29	29						x					
30	12						x					
31	13						x					
32	32											x
33	33											
34	34											
35	35											
36	6					x						

Sobre la base de los resultados, los autores de este escrito consideran dos criterios que pueden tomarse para la eliminación de ítems. El primero de ellos es eliminar reactivos cuyas proporciones de respuesta correcta para cada nivel del rasgo no sean las esperadas por el modelo logístico. Para estar seguros de que esto sucede, dicho reactivo no debería ajustar al modelo en más de una matriz de datos. Esto último es muy importante, ya que si la decisión estadística sobre el ajuste se modifica al replicar el estudio, entonces los resultados presentan ambigüedad y, por lo tanto, no puede tomarse ninguna decisión real fundada en los datos.

El segundo criterio implica proponer un número mínimo de atributos desfavorables para justificar la decisión de eliminar un ítem, tomando como criterio previamente la cantidad total *k* de atributos desfavorables que el investigador desea considerar. Los autores de este trabajo toman como criterio que un ítem es defectuoso si su cantidad de atributos desfavorables es mayor o igual a *k*/2. También es necesario tomar en cuenta que, si la gran mayoría de los ítems posee un número desmedido de atributos desfavorables, entonces conviene mejor rearmar el test.

Vale aclarar que, si los ítems del test se modificaron frente a estudios sucesivos, entonces es necesario considerar qué versión del test se desea mantener definitivamente. En tal caso, es preferible estudiar el segundo criterio de eliminación de ítems en aquella versión del test que el psicómetra desea mantener, ya que se busca mejorar la calidad de medida de ese test y no la de otras variantes del mismo. En el caso del TAF, los autores desean mantener la segunda versión, puesto que la misma involucra avances respecto de la versión piloto. Por lo tanto, las decisiones sobre la eliminación de ítems deberían tomar como referencia principalmente a dicha versión.

Observando la Tabla 2, sólo el ítem 11 no ajusta al ML3P en más de una matriz de datos. Por lo tanto, esto sería suficiente para pensar en su eliminación. Por otra parte, puede apreciarse que, de los seis atributos desfavorables que los autores cuentan para M2, sólo un reactivo cumple con un número mayor o igual que tres; tal reactivo es el 5, puesto que tiene cuatro atributos desfavorables. Por consiguiente, esto permitiría considerar también su eliminación. Los ítems restantes no presentan evidencia de falta de ajuste en más de una matriz de datos, ni poseen tres o más atributos desfavorables en M2.

Discusión

Este manuscrito presenta criterios para la eliminación de ítems desde el marco complementario que brindan la TCT y la TRI. Para ello, se presentan los resultados principales arrojados sobre un Test de Analogías Figurales (TAF) en sucesivas muestras. El presente estudio lleva por ahora a la decisión de eliminar a dos ítems de un total de 36.

Al eliminar sólo dos ítems, la información que brinda el TAF sobre el Razonamiento Analógico se verá poco afectada y, por otro lado, mejorará su calidad psicométrica.

Cabe señalar como limitación del presente estudio, el hecho de que las muestras no fueron obtenidas de manera sistemática, lo cual puede afectar la validez externa de los resultados. Sin embargo, en estudios futuros podría continuarse este análisis psicométrico para verificar los datos aquí presentados, así como comprobar si otros reactivos del TAF deberán ser eliminados o modificados en su contenido.

Referencias

Allen, M. J. & Yen, W. M. (1979). *Introduction to Measurement Theory*. Monterrey, CA: Brooks/Cole Publishing Company.

Argibay, J. C. (2006). Técnicas psicométricas. Cuestiones de validez y confiabilidad. *Subjetividad y procesos cognitivos, 8,* 15-33.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. En F. M. Lord & M. R. Novick (Eds.). *Statistical Theories of Mental Test Scores* (pp. 397-479). Reading, MA: Addison Wesley.

Blum, G. D., Abal, F. J. P., Galibert, M. S., & Attorresi, H. F. (2011). Construcción de una Prueba de Analogias Figurales. *Summa Psicológica UST, 18*(1), 5-12.

Blum, G. D., Galibert, M. S., Abal, F. J. P., Lozzia, G. S. & Attorresi, H.F. (2011). Modelización de una Prueba de Analogías Figurales con la Teoría de Respuesta al Ítem. *Escritos de Psicología, 4*(3), 36-43.

Borg, W. R. (1963). *Educational research: an introduction*. Londres: Longman.

Brown, L., Sherbenou, R. J. & Johnsen, S. K. (2000). *TONI 2. Test de Inteligencia No Verbal. Apreciación de la habilidad cognitiva sin influencia del lenguaje*. Madrid: TEA Ediciones.

Cohen, L. & Manion, L. (2002). *Métodos de investigación cuantitativa*. Madrid: La Muralla.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334.

Embretson, S. E. & Reise, S. P. (2000). *Item Response Theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates Inc.

García-Cueto, E., & Fidalgo, A. M. (2005). Análisis de los ítems. En J. Muñiz, A. M. Fidalgo, E. García-Cueto, R. Martínez & R. Moreno (Eds.), *Análisis de los ítems* (pp. 53-130). Madrid: La Muralla.

Gulliksen, H. O. (1950). *Theory of Mental Tests.* New York: Wiley.
Hambleton R. K, Swaminathan, H. & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Londres: Sage.
Lord, F. M. (1952) A theory of test scores. *Psychometric Monograph Nº7*. Iowa City, IA: Psychometric Society.
Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.
Martínez Arias, M. R., Hernández Lloreda, M. V. & Hernández Lloreda, M. J. (2006). *Psicometría*. Madrid: Alianza.
McDonald, R. P. (1999). *Test theory. A unified treatment.* Mahwah, NJ: Lawrence Erlbaum.
Muñiz, J. (1996). *Psicometría*. Madrid: Universitas.
Muñiz, J. (1997). *Introducción a la teoría de respuesta a los ítems*. Madrid: Pirámide.
Muñiz, J. (2010). Las teorías de los tests: teoría clásica y teoría de respuesta a los ítems. *Papeles del psicólogo, 31*(1), 57-66.
Prieto, G. & Delgado, A. R. (1996). Construcción de los ítems. En J. Muñiz (Coord.), *Psicometría* (pp. 105-138). Madrid: Universitas.
Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhague: The Danish Institute for Educational Research.
Raven, J. C., Court, J. H. & Raven, J. (1993). *Test de matrices progresivas. Escalas coloreada, general y avanzada*. Buenos Aires: Paidós.
Raven, J., Raven, J. C. & Court, J. H. (1991). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Sections 1, 2, 3, & 4*. Oxford: Oxford Psychologists Press.
Santisteban, C. (1990). *Psicometría. Teoría y práctica en la construcción de tests*. Madrid: Norma.
Zimowski, M., Muraki, E., Mislevy, R. & Bock, R. (1996). *BILOG-MGTM: Multiple-group IRT analysis and test maintenance for binary items* [Software]. Chicago, IL: Scientific Software International.