Calibración del Índice de Hiperactividad de Conners mediante el modelo de Rasch

Calibration of Conners ADHD Index with Rasch Model

Recibido: enero 8 de 2012 | Revisado: julio 10 de 2012 | Aceptado: diciembre 26 de 2012

Benito Arias Martínez*
Universidad de Valladolid, España
Víctor B. Arias González**
Universidad de Talca, Chile
Laura Elísabet Gómez Sánchez***
Universidad de Oviedo, España

Resumen

El objetivo del presente estudio se ha centrado en la calibración de la versión española de la escala reducida de 10 ítems (Índice de Hiperactividad de Conners [IHC]) mediante el modelo de Rasch, utilizando una muestra de 482 niños de 5 y 6 años. Los resultados mostraron que el IHC tiene buenas propiedades psicométricas. El ajuste, tanto de los ítems como de las personas al modelo de Rasch, es bueno y las categorías de respuesta funcionan adecuadamente. Las puntuaciones de los niños fueron significativamente superiores a las de las niñas. Tres de los ítems tienen riesgo de presentar Funcionamiento Diferencial del Ítem (DIF) en función del género. El IHC tiene un evidente efecto suelo que impide evaluar a niños con niveles de hiperactividad bajos.

Palabras clave autores

TDAH, calibración de instrumentos, propiedades psicométricas, modelo de Rasch, teoría de respuesta a los ítems, IHC de Conners, modelo logístico de un parámetro.

Palabras clave descriptores

Psicometría, psicología cuantitativa, psicología clínica.

Abstract

The aim of this study was focused on the calibration of the Spanish version of the 10-item Conners Hyperactivity Index or Conners 3-AI using the Rasch model. The participants in this study were 482 children aged 5 and 6 years old. Results showed that the IHC has good psychometric properties. The fit of both items and persons data to Rasch model was good, and the response categories were functioning properly. Boys scored significantly higher than girls. Three of the items were at risk of gender-specific DIF. The Conners 3-AI has an obvious floor effect, which prevents the evaluation of children with low levels of hyperactivity.

Key words authors

Assessment, ADHD, instrument calibration, psychometric properties, Rasch model, item response theory, Conners 3-AI, one-parameter logistic model, 1-PLM.

Key words plus

Psychometry, Quantitative Psychology, Clinical Psychology.

doi:10.11144/Javeriana.UPSY12-3.cihc

Para citar este artículo: Arias, B., Arias, V. B. & Gómez, L. E. (2013). Calibración del Índice de Hiperactividad de Conners mediante el modelo de Rasch. *Universitas Psychologica, 12*(3), 957-970. doi:10.11144/Javeriana.UPSY12-3.cihc

* Universidad de Valladolid, España. Profesor Titular. Departamento de Psicología, ResearcherID: D-7925-2013. E-mail: barias@psi.uva.es

** Universidad de Talca, Chile. Profesor Asistente. Facultad de Psicología, ResearcherID: H-6294-2013. E-mail: viarias@utalca.cl

*** Universidad de Oviedo, España. Doctora. Profesora Ayudante. Departamento de Psicología, ResearcherID: B-5013-2011. E-mail: gomezlaura@uniovi.es

El Trastorno por Déficit de Atención con Hiperactividad (TDAH) es un trastorno neurobiológico de aparición en la infancia caracterizado por síntomas de desatención que puede ir acompañado de hiperactividad y/o impulsividad. Es una de las alteraciones más comunes del comportamiento infantil, cifrándose su prevalencia en tasas que van del 3 % al 7 % (American Psychiatric Association [APA], 2000) en poblaciones no clínicas. No obstante, los datos sobre prevalencia en la población de 6 a 14 años varían notablemente entre los diferentes estudios, dependiendo de los criterios diagnósticos utilizados, del tipo de instrumento de evaluación, del tipo de muestra, de quiénes sean los informantes y de otras variables étnicas, culturales, sociales o geográficas (Amador, Forns & Martorell, 2001a, 2001b; Barkley & Murphy, 2006; Cardo & Servera, 2005; Reid et al., 1998).

Los métodos de evaluación del TDAH más usados han sido las escalas de clasificación cumplimentadas por padres y maestros (Barkley, 1990; Conners, 1997; DuPaul, Power, Anastopoulos & Reid, 1998). Muchas de ellas se han desarrollado tomando como base la sintomatología descrita por el *DSM-IV-TR* (Barkley, 2006; Burns, Walsh & Gomez, 2003; DuPaul et al., 1998; Gadow & Sprafkin, 1997; Gomez, 2007; Swanson, 2010; Wolraich et al., 2003). Junto con dichos instrumentos, las escalas de Conners (Conners, 1997) son, posiblemente, los instrumentos más utilizados en los últimos años en la evaluación del TDAH. Se trata de cinco escalas probabilísticas de estimaciones sumatorias con cuatro opciones de respuesta desde *nada* (0) a *mucho* (3) que pueden aplicarse de forma independiente. Las dos primeras (Conners 3–P y Conners 3–T) son cumplimentadas, respectivamente, por padres y maestros y van dirigidas a evaluar niños y adolescentes de 6 a 18 años. La tercera (Conners 3–SR) tiene formato de autoinforme y se aplica, junto con las dos anteriores, a individuos que tengan entre 8 y 18 años de edad. La cuarta (Conners 3 Global Index o Conners 3–GI) es una medida de la psicopatología general del niño. Por último, la Conners 3 ADHD Index o Conners 3–AI) consta de 10 ítems considerados prototípicos de la hiperactividad (de hecho, son aquellos que en los análisis factoriales de las escalas para padres y profesores han presentado saturaciones mayores). Las escalas Conners han mostrado propiedades psicométricas adecuadas de fiabilidad y validez en muestras estadounidenses (Barkley, 2006; Conners, 1997). El rango de los coeficientes de consistencia interna va de 0.77 a 0.97; y los coeficientes de fiabilidad test-retest van de 0.71 a 0.98; los coeficientes de concordancia entre evaluadores se sitúan entre 0.52 y 0.94. Las evidencias de validez se han establecido mediante análisis factorial exploratorio y confirmatorio, y se ha determinado asimismo que las escalas cuentan con suficientes evidencias de validez de constructo (mediante la comparación con otras medidas) y predictiva (las escalas se han mostrado eficaces para discriminar entre sujetos diagnosticados y no diagnosticados de TDAH).

Se quiere poner de manifiesto, por una parte, la importancia que tanto para la investigación como para la práctica clínica supone contar con instrumentos de evaluación cuyas características psicométricas se hayan determinado con precisión y, por otra, las ventajas que la Teoría de Respuesta a los Ítems (TRI) presenta sobre la Teoría Clásica de los Test (TCT). Con base en esas premisas, el objetivo del presente estudio se ha centrado en la calibración de la adaptación española de la escala reducida de 10 ítems Conners 3 ADHD Index o Conners 3–AI (en español Índice de Hiperactividad de Conners o IHC) mediante el modelo de Rasch o modelo logístico de un parámetro, 1P-LM (encuadrado habitualmente en la TRI), toda vez que hasta el momento el acercamiento más común al análisis de las escalas de Conners ha sido la TCT (Amador, Idiazábal, Aznar & Peró, 2003; Farré-Riba & Narbona, 1997) con alguna excepción (Gumpel, Wilson & Shalev, 1998).

Las ventajas de la TRI (y del modelo de Rasch) respecto a la TCT han sido profusamente difundidas (Andrich, 1988; Ayala, 2009; Bond & Fox, 2001; Crocker & Algina, 2008; Embretson & Hershberger, 1999; Embretson & McCollam, 2000; Embretson & Reise, 2000; Fidalgo, 2005; Hambleton, Swaminathan & Rogers, 1991; Prieto & Delgado, 1999, 2000; Prieto & Dias, 2004; Wright & Stone,

1979). Las más relevantes son: invarianza de los parámetros en distintas muestras; posibilidad de realizar estimaciones del grado de precisión con la que cada test (y cada ítem individual) mide los diferentes niveles de habilidad de los sujetos examinados; independencia de la estimación de *theta* (θ) respecto a la prueba utilizada; medición conjunta (los parámetros de las personas y de los ítems se expresan en las mismas unidades intervalares (*logits*) y se localizan en el mismo continuo); objetividad específica (la diferencia entre dos personas en un atributo no depende de los ítems específicos con los que sea estimada); propiedades de intervalo (la interpretación de las diferencias en la escala es la misma a lo largo del atributo medido, de modo que a diferencias iguales entre un sujeto y un ítem le corresponden probabilidades idénticas de dar una respuesta correcta); especificidad del error típico de medida y, finalmente, facilidad para la personalización de las pruebas.

Método

Participantes

Participaron en este estudio 551 alumnos de Educación Infantil (264 niñas y 287 niños), seleccionados de modo incidental en 14 centros escolares de dos ciudades de la Comunidad Autónoma de Castilla y León, España. La media de edad de los 482 casos (excluyendo los valores perdidos) en esta variable fue de 5.3 años (DE = 0.99) en el caso de las niñas y de 5.23 años (DE = 0.96) en el caso de los niños. La distribución de niños (N = 254) y niñas (N = 228) en función de la edad en meses presenta unos valores moderados tanto en asimetría (A_s = -0.005, ET = 0.153) como en curtosis (g_2 = -0.947, EE = 0.304). El único criterio de selección utilizado fue que los niños cursaran uno de los dos cursos de Educación Infantil. Se trata, por tanto, de una muestra comunitaria (i. e., no clínica) seleccionada en función de la disponibilidad de acceso a los sujetos participantes.

Se sometieron a prueba distintas hipótesis de equiprobabilidad, tomando en consideración las variables curso, género y centro escolar al que asistían los alumnos. Los resultados del análisis de residuos estandarizados de Pearson pusieron de manifiesto que es plausible aceptar la hipótesis de independencia entre esas variables. Así, el contraste para (CURSO) ha sido $\chi^2_{(1)} = 0.744$; $p = 0.388$; para {CURSO}{GÉNERO}, $\chi^2_{(1)} = 0.293$; $p = 0.588$ y para [CURSO, GÉNERO][CENTRO] $=\chi^2_{(39)} = 36.126$; $p = 0.602$.

Instrumento

Se utilizó como instrumento de medida el Índice global de Hiperactividad de Conners (IHC, Conners 3–AI), compuesto por 10 ítems (Tabla 1), adaptado *ad hoc* para la presente investigación. En este estudio se asumió la hipótesis –y los datos así lo corroborarán, como el lector podrá comprobar en la sección de Resultados– de que su estructura es unidimensional.

Esta escala, en su versión original, ha demostrado repetidamente su capacidad para discriminar entre sujetos con y sin síntomas clínicos de TDAH, y su validez de constructo se ha establecido mediante análisis factorial en muestras estadounidenses (Conners, 1989, 1997; Merrell, 2003). Los estudios sobre las cualidades psicométricas de la escala original señalan que la consistencia interna (evaluada mediante el coeficiente alfa de Cronbach) es adecuada, oscilando entre 0.73 y 0.94 en las versiones de padres; entre 0.77 y 0.95 en las versiones para profesores y entre 0.75 y 0.92 en el caso del autoinforme. Finalmente, la estabilidad temporal con administración de la prueba en intervalos de 6 y 8 semanas ha resultado aceptable, con valores de r entre 0.47 y 0.88 (Merrell, 2003).

Las versiones abreviadas de las escalas de Conners y en especial la Escala Abreviada TADH de 10 ítems o IHC (su contenido puede consultarse en la Tabla 1), cuyo análisis es el objeto de este trabajo por su facilidad y rapidez de aplicación, están indicadas para su uso como instrumentos de investigación con muestras amplias, así como de cribado o *screening* en procesos de evaluación que abarquen cantidades amplias de sujetos. El formato de respuesta es una escala de cuatro puntos donde 0 = *nada*; 1 = *algo*; 3 = *bastante* y 4 = *mucho*.

Procedimiento

Se confeccionó una versión electrónica del IHC mediante el programa LimeSurvey v. 1.87 (LimeSurvey, 2009) con el objeto de que se pudiera cumplimentar por internet. A continuación, se contactó a las profesoras (25 en total) de los centros escolares y se les proporcionó la información necesaria sobre la investigación, junto con un modelo de consentimiento informado que debían aceptar los padres o representantes legales de los alumnos. Se les dio un plazo de dos semanas para cumplimentar la escala, pasado el cual se recuperó la información de la base de datos y se procedió a su análisis.

Análisis de datos

En la presente investigación se utilizó el Modelo de Escalas de Clasificación de Rasch (Rasch Rating Scale Model [RSM]; Andrich, 1978; Rasch, 1960, 1977; Wright & Masters, 1982) implementado en el programa WINSTEPS v. 3.73 (Linacre, 2011; Linacre & Wright, 1999). Se empleó, para realizar otros análisis, el programa SAS v. 9.2.

El RSM especifica la probabilidad, P_{nij}, de que una persona *n* con un nivel de habilidad χ_n sea observada en una categoría *j* de una escala de clasificación aplicada al ítem *i* de un nivel de dificultad (o 'probabilidad de adhesión') χ_i como opuesto a la probabilidad $P_{ni(j-1)}$ de ser observada en una categoría *(j-1)*. Así, en una escala Likert, *j* podría ser *mucho* y *j-1* podría ser *bastante*.

$$\log_e\left(P_{nij} \middle/ P_{ni(j-1)}\right) = \beta_n - \delta_i - \tau_j \qquad (1)$$

En la ecuación (1) T_j es el "umbral de Rasch-Andrich" (Rasch-Andrich threshold), también denominado "calibración de paso" (*step calibration*) o "dificultad de paso" (*step difficulty*). El modelo resulta apropiado para estimar la variable latente (hiperactividad) y la probabilidad de adhesión de los ítems para respuestas puntuadas en dos o más categorías; además, asume que la distancia entre los parámetros de umbral es constante a través de todos los ítems (Ayala, 2009; Linacre, 2011).

Resultados

Evaluación del ajuste de los ítems y las personas

En la Tabla 1 se ofrecen los resultados del análisis de Rasch, ordenados en función de la magnitud de los parámetros estimados. La polaridad de los ítems indica que todas las correlaciones punto-biseriales resultaron positivas y superiores al valor recomendado de 0.2 (Linacre, 2011), con un rango de 0.52 a 0.76. En consecuencia, todos los ítems cumplen el requisito crucial en el análisis Rasch de estar alineados en la misma dirección en la variable latente.

El índice de separación de los ítems fue de 7.21, indicando que estos discriminan entre diferentes niveles de hiperactividad entre los sujetos. La fiabilidad global o Item Separation Reliability (0.98) indica que los ítems configuran una variable que está bien definida y que la fiabilidad de la ubicación de los ítems en la escala es buena, aportando asimismo evidencia del cumplimiento del supuesto de independencia local. Con esta muestra, las dificultades de los ítems se estiman con mucha exactitud.

Las estimaciones de los sujetos también resultaron fiables. La separación es de 2.02. Este índice evalúa en qué medida el test discrimina a la muestra en suficientes niveles para el propósito del estudio. Equivale, aproximadamente, a un valor KR-20 o alfa de Cronbach de 0.8, y denota que en la muestra estudiada el IHC discrimina al menos entre dos niveles (i. e., sujetos con hiperactividad baja y alta). El índice de fiabilidad promedio de las personas (Person Separation Reliability) es apropiado (0.8).

Los valores RS-MC (Raw Score-to-Measure Correlation) son las correlaciones de Pearson entre las puntuaciones directas y las medidas (i. e., parámetros estimados por el modelo), incluyendo las puntuaciones extremas. Se espera que, cuando los datos son completos, sean cercanos a 1 para las personas y a -1 para los ítems (lo que se cumple en el presente caso, puesto que se han obtenido 0.96 y -1, respectivamente).

El ajuste medio y las desviaciones estándar de los ítems son apropiados (*Infit* = 1.01; *DE* = 0.25; *Outfit* = 0.98; *DE* = 0.28). El ajuste medio

Tabla 1
Estimaciones de los parámetros de los ítems

Ítem		MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZEMP	OUTFIT MNSQ	OUTFIT ZEMP	PTBISERL-EX CORR.	PTBISERL-EX EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	ESTIM DISCR
8 (llora con facilidad)		0.87	0.09	1.48	1.7	1.51	1.4	0.42	0.61	63.6	68.4	0.51
9 (cambios estado ánimo)		0.78	0.09	0.83	-0.8	0.68	-1.1	0.68	0.62	73.1	67.4	1.22
10 (rabietas)		0.63	0.09	0.81	-0.9	0.75	-0.9	0.72	0.63	74.5	67.1	1.25
5 (agitado nervioso)		0.42	0.09	0.72	-1.3	0.68	-1.3	0.78	0.64	71.9	66	1.32
3 (molesta interrumpe)		0.09	0.09	0.82	-0.8	0.75	-1.1	0.76	0.66	71.4	64.2	1.25
7 (frustración fácil)		0.01	0.09	0.98	-0.1	0.97	-0.1	0.63	0.66	62.9	63.9	0.99
4 (dif. terminar tareas)		-0.33	0.08	1.43	1.7	1.42	1.6	0.53	0.67	53.3	62	0.54
2 (excitable impulsivo)		-0.4	0.08	0.88	-0.5	0.88	-0.5	0.74	0.68	64.3	61.7	1.14
6 (desatento se distrae)		-0.94	0.08	1.05	0.3	1.05	0.2	0.62	0.69	56	58.2	0.91
1 (inquieto sobreactivo)		-1.13	0.08	1.05	0.3	1.06	0.3	0.69	0.69	55.7	57.4	0.92
ITEMS	MEDIA	0	0.09	1.01	0	0.98	-0.2			64.7	63.6	
	DE	0.66	0.01	0.25	1	0.28	1			7.4	3.6	
PERSONS	MEDIA	-2.3	0.77	1	-0.1	0.97	-0.1			64.7	63.6	
	DE	2.07	0.45	0.61	1	0.64	1			21.1	11.6	
Item separation		7.21	Item reliability		0.98	RMSE	0.09	RS-MC	-1			
Person separation		2.02	Person reliability		0.8	RMSE	0.92	RS-MC	0.96	Cronbach alpha		0.9

Nota. MEASURE = Estimación (calibración) del parámetro; MODEL S.E. = Error estándar de la estimación; INFIT y OUTFIT MNSQ = Estadísticos estandarizados de información ponderada por cuadrados medios; PTBISERL-EX (CORR. y EXP.) = Correlación punto-biserial (observada y esperada por el modelo); EXACT MATCH (OBS% y EXP%) = Porcentaje de puntos que se ajustan a la predicción y porcentaje esperado por el modelo; ESTIM DISCR = Discriminación estimada; MEAN = Media; *DE* = Desviación Estándar; RMSE = Root-mean-square average of the standard errors (Promedio de raíz cuadrática media de los errores estándar); RS-MC = Raw Score to Measure Correlation (Correlación entre puntuaciones directas y parámetros estimados por el modelo); Separation = Índice de Separación; Reliability = Índice de Fiabilidad.
Fuente: elaboración propia.

y las desviaciones estándar de los sujetos son asimismo apropiados (*Infit* = 1; *DE* = 0.61; *Outfit* = 0.97; *DE* = 0.64). Estos resultados sugieren que este conjunto de ítems cumple en principio los requisitos necesarios para identificar síntomas hiperactivos.

Una representación gráfica del ajuste mediante *Infit* y *Outfit MNSQ* puede verse en la Figura 1. Los dos ítems más fáciles (i. e., con más probabilidades de adhesión) son el 1 y el 6, en tanto que los de adhesión más difícil son el 8 y el 9. Todos los errores estándar son parejos y razonablemente reducidos. Finalmente, todos los ítems se sitúan en la zona de 0.5 a 1.5, o zona de ajuste aceptable y productivo para la medida (Linacre, 2011), con la única excepción del ítem número 8, cuyo valor *Outfit MNSQ* ha presentado una desviación mínima del rango ideal (1.51). En cuanto a los sujetos, también se constata un adecuado ajuste al modelo: el *Infit* promedio ha sido 1 (*DE* = 0.61) y el *Outfit* promedio 0.97 (*DE* = 0.64). Cabe señalar que únicamente el 12.86 % y el 14.31 % presentaron valores superiores a 1.5 en *Infit* y *Outfit MNSQ*, respectivamente. En consecuencia, asciende a un 87.14 % (o al 85.68 % si se considera también el *Outfit*) la proporción de alumnos que presentan un buen ajuste.

Objetividad específica

El análisis de la objetividad específica se ha llevado a cabo dividiendo la muestra original en dos submuestras aleatorias de 241 sujetos cada una y realizando a continuación un análisis de regresión lineal simple entre los parámetros de dificultad de los ítems obtenidos en cada una de ellas (Hamble-

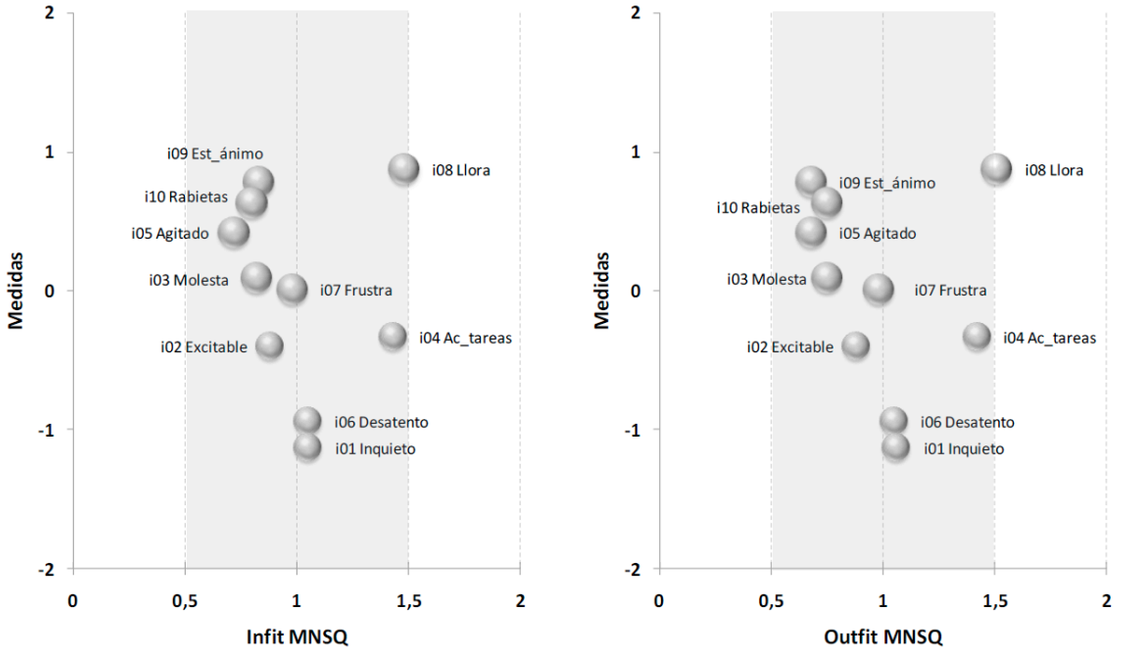

Figura 1. Distribución de los ítems en función de los valores Infit y Outfit MNSQ.
Nota. El diámetro de las burbujas corresponde al error estándar, el área sombreada indica la zona de ajuste óptimo.
Fuente: elaboración propia.

ton et al., 1991; Prieto & Delgado, 2003). La correlación entre ambos conjuntos de parámetros ha sido 0.979, la intercepción -0.001, la pendiente 1.108 y el coeficiente de determinación 0.959. Puesto que los valores que denotarían un ajuste perfecto entre los datos y el modelo serían 1, 0, 1 y 1, respectivamente, se concluyó que se cumple el requisito de invarianza de los parámetros de los ítems y que los datos presentan un buen ajuste global al modelo.

Adecuación del nivel de dificultad de los ítems para la muestra

Los mapas de personas e ítems, también conocidos como "Mapas de Wright", ilustran gráficamente cómo los ítems progresivamente mayores en nivel de dificultad se van solapando con los niveles de las personas en el rasgo latente (hiperactividad) evaluado. Dado que el modelo de Rasch utiliza la misma medida (*logit*) ambas métricas pueden compararse para determinar si la dificultad de los ítems es o no apropiada para la muestra de alumnos. Si la muestra fuera apropiada, debería existir un solapamiento considerable en el mapa entre los parámetros de dificultad de los ítems y los niveles del rasgo latente de las personas. A este alineamiento entre ítems y personas se le denomina *targeting* en el argot del análisis Rasch.

En la Figura 2 se muestra el mapa conjunto de ítems y personas ordenados desde los niveles más altos a los más bajos. Por consiguiente, los alumnos con niveles elevados de hiperactividad, así como los ítems que miden niveles de hiperactividad más severos, se encuentran en la parte alta del mapa. Se puede comprobar cómo el rango de los parámetros de dificultad de los ítems se solapa parcialmente con el rango de los parámetros del rasgo latente de los alumnos, lo que indica que los 10 ítems han evaluado a los sujetos con distintos niveles de hiperactividad.

Con todo, es preciso hacer las consideraciones siguientes: En primer lugar, la media de las personas ($M = -2.3$; $DE = 2.07$) es muy inferior a la de los ítems ($M = 0$; $DE = 0.66$). En segundo lugar, la amplitud

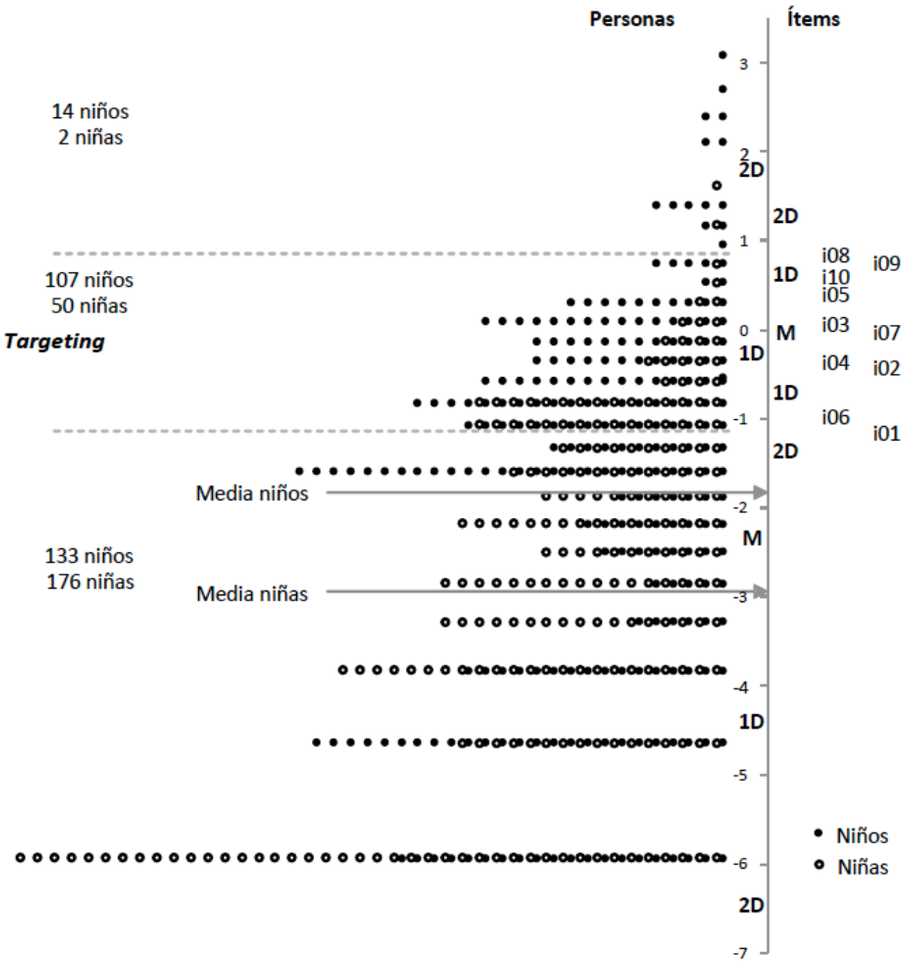

Figura 2. Mapa de Wright (personas-ítems).

M = Media; *1D* = Desviación Estándar; *2D* = Dos desviaciones Estándar. i01 = Inquieto o sobreactivo; i02 = Excitable, impulsivo; i03 = Molesta o interrumpe a otros niños; i04 = Tiene dificultad para acabar las tareas que empieza; i05 = Agitado, nervioso; i06 = Desatento, se distrae con facilidad; i07 = Sus demandas deben satisfacerse inmediatamente (se frustra con facilidad); i08 = Llora con facilidad; i09 = Experimenta cambios drásticos y súbitos en su estado de ánimo; i10 = Rabietas, conducta explosiva e impredecible.

Fuente: elaboración propia.

de las personas (de -5.93 a 3.1 *logits*) es muy superior a la de los ítems (de -1.13 a 0.87 *logits*). En tercer lugar, un total de 309 sujetos (64.1 %) puntúan por debajo del rango de dificultad de los ítems, en tanto que solo 16 (3.3 %) lo hacen por encima de dicho rango. La zona de *targeting* o alineamiento entre la dificultad de los ítems y la presencia del rasgo latente en los sujetos agrupa a 157 niños (32.6 %). En cuarto lugar, se advierten notorias diferencias entre el número de niños y niñas en cada una de las tres zonas mencionadas. Un análisis de los residuos estandarizados de Pearson (Figura 3) nos lleva a la conclusión de que las distribuciones por género en las diferentes zonas (*targeting*, superior e inferior), son significativamente distintas a las esperadas bajo el supuesto de equiprobabilidad ($\chi^2_{(2)} = 35.9$; $p < 0.001$).

Así, el número de mujeres en la zona inferior es mayor que el esperado bajo el supuesto de independencia entre las variables género y zona de calificación, y lo contrario sucede en el caso de los

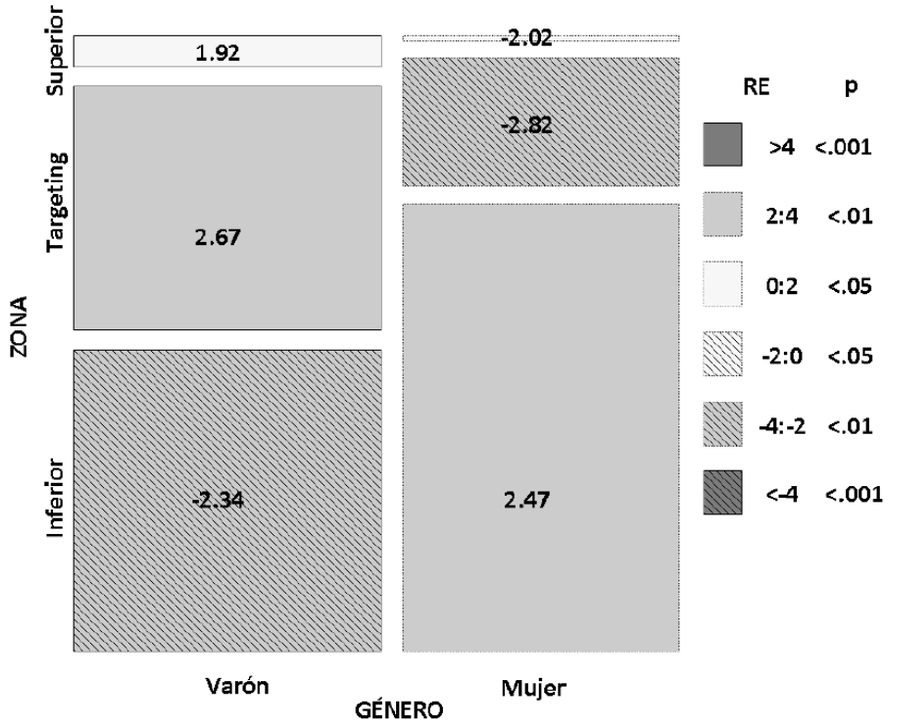

Figura 3. Residuos estandarizados de Pearson por género y zona de calificación en el IHC.
RE = Residuos estandarizados de Pearson; *p* = nivel de significación a posteriori.
Fuente: elaboración propia.

varones. En la zona de *targeting*, el número de varones es superior al esperado por el modelo, y el de mujeres inferior al esperado. En la zona superior, los residuos de los varones son positivos ($p > 0.05$) y los de las mujeres negativos ($p < 0.05$). Estas diferencias se ven refrendadas por el análisis de varianza. Las medias de niñas y niños fueron de -2.93 ($DE = 1.97$) y -1.81 ($DE = 2.07$), respectivamente ($F_{(1,480)} = 37.77$; $p < 0.001$; $d = 0.56$), lo que denota una diferencia significativa de tamaño medio (Cohen, 1988) a favor de los niños en las puntuaciones del IHC.

Evaluación de la dimensionalidad

Una de las asunciones subyacentes al modelo de Rasch es que la escala sea unidimensional. Se comprobó este requisito mediante los estadísticos de ajuste *Infit* y *Outfit* consignados en párrafos anteriores y mediante el análisis de componentes principales (ACP) de los residuos estandarizados de Rasch. El ACP descompone la matriz de correlaciones entre los ítems basándose en los residuos estandarizados (i. e., diferencias entre los valores observados y los predichos por el modelo de Rasch) para determinar si existen o no otras dimensiones potenciales.

El primer factor del análisis corresponde a la Dimensión Rasch. Se considera adecuada una varianza igual o superior al 60 %. La segunda dimensión (o primer contraste de los residuos) indica si existen patrones en las diferencias dentro de los residuos suficientemente grandes como para sugerir que es plausible la existencia de más de una dimensión. El análisis de componentes principales del IHC muestra que el 51 % de la varianza queda explicada por los datos modelados. Este porcentaje es algo inferior al recomendado del 60 %. El primer contraste presenta un valor propio de 2.6 (inferior al valor de 3 necesario para considerar una segunda dimensión), lo que indica que contiene menos de 3 ítems, y explica el 12.6 % de la varianza de los datos no modelados. En consecuencia con lo dicho,

la escala podría considerarse unidimensional (o "suficientemente unidimensional" si se quiere ser más preciso).

Función de las categorías de respuesta

Se revisaron a continuación las clasificaciones de cada ítem a fin de determinar si las categorías de respuesta funcionaban según lo esperado. En primer lugar, todas las frecuencias de las cuatro categorías utilizadas (*nada*, *algo*, *bastante*, *mucho*) excedieron el mínimo de 10 recomendado por Linacre (1999). La más frecuente fue *nada* (N = 2282) seguida de *algo* (N = 1733), *bastante* (N = 651) y *mucho* (N = 153). Los valores *Infit MNSQ* fueron próximos al valor esperado de 1 en todas las categorías (1.03, 0.94, 1 y 1.1, respectivamente). Los valores *Outfit MNSQ* también fueron próximos a 1 en las cuatro categorías (1.02, 0.87, 1 y 1.13, respectivamente), lo que indica que la categoría proporciona más información (i. e., varianza sistemática) que ruido (i. e., varianza de error) en el proceso de medición (Linacre, 1999).

En segundo lugar, se constató que todas las medidas promedio para las categorías avanzaran monotónicamente (i. e., *nada* → *algo* → *bastante* → *mucho*), y no hubiera ninguna categoría especialmente ruidosa. Así, las medidas promedio (-3.14; -1.34; -0.09 y 1.13) y las estimaciones de los umbrales (-2.24; 0.27 y 1.97) presentan un incremento que corre parejo con la progresión de las etiquetas de las categorías, lo que sugiere que la categorización de la escala de clasificación ha sido satisfactoria (Figura 4). Como es obvio, los umbrales de paso (τ_n) entre las categorías son solo tres, dado que las categorías son cuatro. La secuencia, por tanto, es la siguiente: $\tau_1 < \tau_2 < \tau_3$, siendo

$$\sum_{m=1}^{4} \tau_m = 0 : -2.24 + 0.27 + 1.97 = 0$$

Esta secuencia de valores está indicando que los parámetros de umbral de Rasch-Andrich están ordenados: -2.24 < 0.27 < 1.97. Junto a los valores de estos parámetros de umbral se muestra el error estándar de los pasos del ítem, observándose que los valores son relativamente bajos (0.04, 0.05 y 0.1).

Al observar el gráfico de las curvas características de las categorías de respuestas (CCCR) podrá apreciarse con más claridad cuál es la categoría de respuesta más probable a lo largo del continuo. Esta curva relaciona la probabilidad de respuesta a un ítem con su nivel en el constructo medido con el test, siendo útil en la evaluación de las propiedades de los ítems. Como puede observarse en la Figura 4, los puntos de intersección entre las categorías de respuestas coinciden con los parámetros de umbral de la medida (**T**). A su vez, estos puntos definen en el continuo las regiones de respuestas más probables.

Análisis del DIF

El análisis del DIF (Funcionamiento Diferencial del Ítem) uniforme en función del género de los participantes ha revelado que en la escala existen tres ítems con cierto riesgo de presentar DIF (en esta muestra). El ítem 5 ("Agitado, nervioso") es 0.71 *logits* más difícil para los niños ($t_{(390)}$ = -3.76; p = 0.0002) que para las niñas. El ítem 7 ("Sus demandas deben satisfacerse inmediatamente - se frustra con facilidad") es 0.58 *logits* más difícil para las niñas ($t_{(404)}$ = -3.4; p = 0.0007), en tanto que el ítem 8 ("Llora con facilidad") es 0.67 *logits* más difícil para las niñas ($t_{(404)}$ = -3.52; p = 0.0005). No se ha apreciado la existencia de DIF no uniforme en ninguno de los 10 ítems.

Precisión de los ítems

En lo que se refiere a la pregunta formulada sobre la precisión de las puntaciones aportadas por los ítems del test, se han estimado las funciones de información de los ítems y la del test global. El resultado que se ha obtenido es que para valores de theta comprendidos entre θ = 0.5 y θ = 1 se observan las puntuaciones de la función de información del test más altas, de ahí que sea en esta región del continuo donde el test mide con una mayor precisión. El mayor error típico de medida se sitúa en las posiciones extremas del continuo (-4.64 y 4.3).

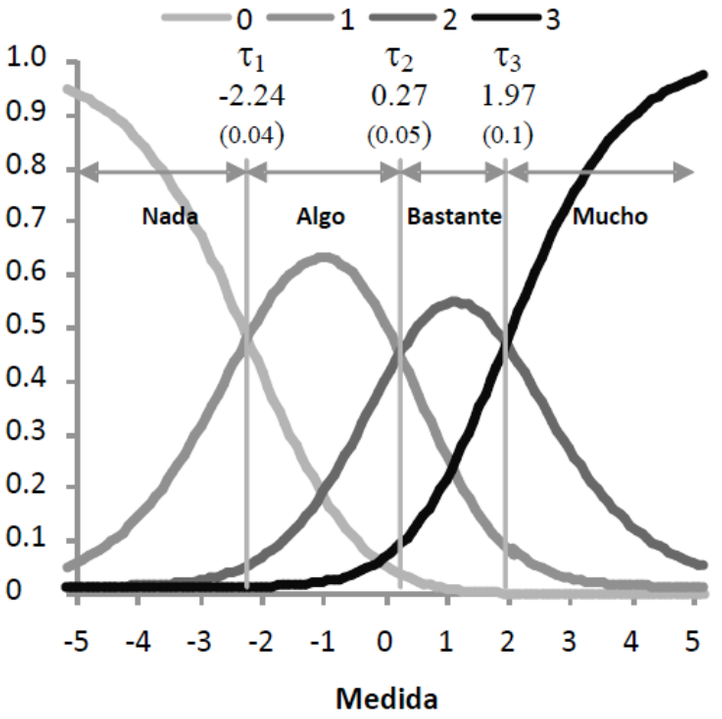

Figura 4. Curvas de probabilidad de las categorías de respuesta.
Nota. Las líneas verticales señalan los umbrales o pasos ('thresholds') de Rasch-Andrich (punto en que dos categorías adyacentes son igualmente probables). Entre paréntesis se señalan los errores estándar.
Fuente: elaboración propia.

Baremos de la prueba

Se construyeron, finalmente, los baremos normalizados del IHC. En ellos se muestran las puntuaciones directas en el test (de 0 a 30), las medidas (i. e., parámetros estimados de Rasch) junto con su error estándar para cada puntuación, las puntuaciones normadas con su error estándar, las frecuencias absolutas y acumuladas y los percentiles. En la Figura 5 se compendia la información de los baremos. A partir de una puntuación directa, se puede estimar con rapidez el resto de valores trazando las correspondientes proyecciones ortogonales contra los distintos ejes de ordenadas en el gráfico. A título de ejemplo, a una puntuación directa de 20 en el IHC (línea vertical punteada) correspondería una medida de Rasch de 1.18, un error estándar de 0.47, una puntuación normada de 668 (*DE* = 23) y un percentil de 97, además de constatar en qué zona del histograma de frecuencias se encuentra la puntuación obtenida.

Discusión

Se ha dedicado este estudio a la calibración del IHC de Conners (Conners 3–AI), utilizando el modelo de Escalas de Clasificación (Rating Scale Model o RSM de Rasch-Andrich). Hasta donde se sabe, el IHC no ha sido –al menos en España– calibrado y validado psicométricamente mediante procedimientos encuadrados en la Teoría de Respuesta a los Ítems; calibración y validación que, indudablemente, tendrán repercusión tanto en la investigación sobre este instrumento como en su aplicación al ámbito clínico. Como se argumentó en la introducción, ese marco metodológico dota a los instrumentos de evaluación de una serie de ventajas que no ofrece la metodología de análisis tradicional (i. e., Teoría Clásica de los Test).

Se han cumplido satisfactoriamente los requisitos de alineación de los ítems con la variable latente, así como la objetividad específica y la independencia local, y se obtuvieron índices de separación de

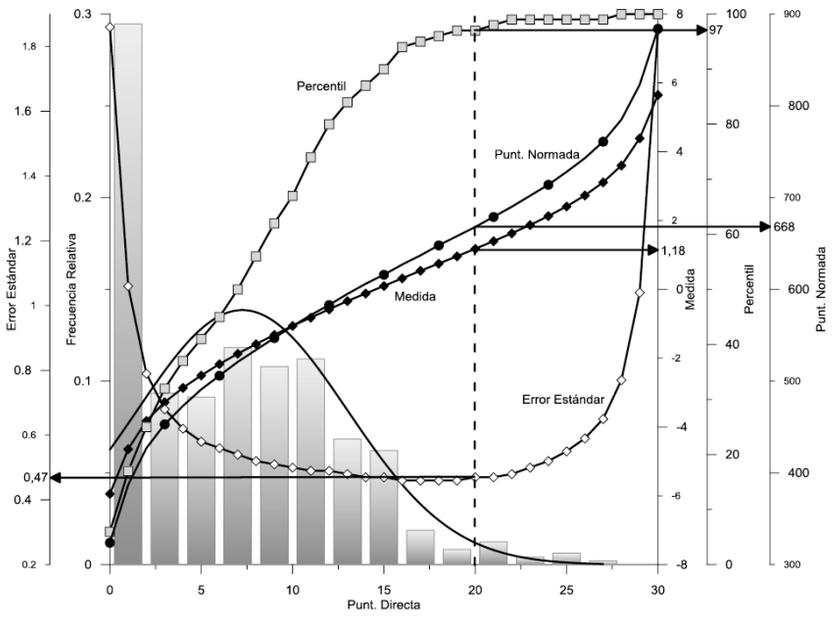

Figura 5. Compendio de la información del análisis Rasch (baremos normalizados).
Fuente: elaboración propia.

los ítems y personas superiores a lo aceptable (0.8). En términos generales, se encontró un buen ajuste de las personas y los ítems al modelo: los ítems del IHC permiten identificar un rango relativamente amplio de síntomas hiperactivos, y tanto el índice de fiabilidad promedio de los ítems como el de las personas y el índice de fiabilidad global han resultado aceptables. Existe un ítem ("Llora con facilidad") que muestra un desajuste mínimo, suficientemente reducido sin embargo como para desechar su eliminación de la Escala. Por lo tanto, se estima que los datos recopilados con los 10 ítems pueden ser explicados convenientemente por el RSM. El ajuste de las personas ha mostrado que para un 85.68 % de los alumnos (considerando conjuntamente los desajustes señalados por *Infit* y *Outfit MNSQ*), la aplicación del RSM al conjunto de ítems del IHC permite explicar convenientemente los patrones de respuestas. Por lo tanto, se puede afirmar que es útil para medir la Hiperactividad en la población a la que se ha administrado la escala, resultado que consideramos de utilidad de cara a la aplicación del instrumento en el ámbito clínico.

Por otra parte, la aplicación del RSM al conjunto de ítems ha permitido conocer otras propiedades de los ítems, tales como el error de medida, las curvas características de las categorías de respuesta y la posición de los ítems en el continuo Hiperactividad. En relación con si los ítems se ordenan homogéneamente y de forma jerárquica respecto a la variable latente evaluada, los ítems han mostrado que se distribuyen a lo largo del continuo, sin excesivos saltos entre ellos, por lo que en principio no resultaría necesario reconstruir el instrumento añadiendo ítems destinados a llenar esos vacíos de información. Los resultados indican por tanto que los ítems se distribuyen de forma jerárquica y con un escalamiento adecuado. En relación con el funcionamiento de las categorías de respuesta y su función de información, resultan adecuadas. Como se ha visto, las CCCR han mostrado que las categorías de respuesta están ordenadas en todos los ítems, tal como exige el modelo (Wright & Masters, 1982). Las categorías demuestran que funcionan según lo esperado, siendo las que aportan mayor información, en este orden: *bastante*, *algo*, *mucho*, *nada*.

La zona de alineamiento de los ítems que componen el IHC se corresponde aproximadamente con un tercio de los sujetos de la muestra. Casi dos tercios se sitúan por debajo del rango de dificultad

de los ítems y únicamente el 3 %, por encima. Esto no es extraño si se considera que se ha aplicado el IHC a una muestra no clínica. Una conclusión obvia de este hallazgo es que, si se desearan evaluar niveles de hiperactividad más bajos, habría que incluir ítems con niveles menores de dificultad de adhesión. Con todo, se estima que aun con ese *efecto suelo* el IHC cumple con su cometido de *screening* y descarta correctamente a los niños sin sospecha de padecer TDAH. En lo que hace referencia a la utilidad clínica, se considera que deberían ser candidatos a una evaluación más exhaustiva aquellos niños que obtuvieran una puntuación directa en el IHC de 20 o superior (correspondiente a 1.18 *logits*, situada en el percentil 97).

En relación con el análisis del funcionamiento diferencial de los ítems, se encontró que, en función del género, en la muestra estudiada existe sospecha de funcionamiento diferencial en los siguientes ítems: 5 ("Agitado, nervioso"), 7 ("Sus demandas deben satisfacerse inmediatamente") y 8 ("Llora con facilidad"). En este sentido, convendría verificar si el DIF se constata también en otras muestras, tanto generales como clínicas o subclínicas. En todo caso, sería excesivo plantear las implicaciones que esa presunta presencia de DIF pudiera tener en la clínica, ya que, por una parte, el nivel de significación asociado a los estadísticos de Mantel-Haenszel no es excesivamente reducido y, por otra, los resultados deben circunscribirse estrictamente a la muestra utilizada.

En lo que se refiere a la precisión de las puntaciones aportadas por los ítems del test, se han estimado las funciones de información de los ítems y la del test global. El resultado que se ha obtenido es que para valores θ comprendidos entre $\theta = 0.5$ y $\theta = 1$ se observan las puntuaciones de la función de información del test más altas, de ahí que sea en esta región del continuo donde el test mide con una mayor precisión. El mayor error típico de medida se sitúa, de acuerdo a lo esperado por el modelo, en las posiciones extremas del continuo (-4.64 y 4.3). En futuras aplicaciones del test y, pensando en la creación de un banco de ítems, el conocer la localización de los ítems en el continuo de Hiperactividad y dónde aporta la máxima información cada ítem, permitiría crear test para los niveles deseados de comportamiento hiperactivo, considerado este como un continuo que iría desde niveles muy bajos de actividad hasta niveles extremadamente elevados, que indicarían la probable presencia de un trastorno de la conducta.

Como limitaciones de la presente investigación, cabe señalar las siguientes. En primer lugar, el carácter incidental de la selección de los sujetos implica que no sea posible la generalización de los resultados a la población. Sería conveniente, para paliar tal inconveniente, utilizar muestras probabilísticas en futuros estudios sobre este problema. En segundo lugar, los resultados han puesto de manifiesto el evidente efecto suelo del IHC, lo que supone que no es una prueba válida para evaluar o detectar niveles de hiperactividad bajos. En tercer lugar, es en cierto modo cuestionable la unidimensionalidad del constructo tal como lo evalúa el IHC en esta muestra. Futuros estudios deberían determinar la existencia de más de una dimensión. En cuarto lugar, cabría señalar algún solapamiento entre los ítems en cuanto a su dificultad (la diferencia en *logits* entre algunos de ellos es muy pequeña). No obstante, se considera que deben mantenerse en la escala, toda vez que su contenido alude a conceptos claramente distintos, si bien todos relacionados con el significado de la variable latente evaluada. Por último, cabría considerar como limitación el modo en que se recogieron los datos (i. e., mediante el uso de un cuestionario cumplimentado por las maestras de los niños); en este sentido, convendría en futuras investigaciones complementar ese método de recogida de información con otros (p. ej., observación directa de la conducta) que permitieran aportar evidencias de la validez criterial del instrumento de evaluación.

Referencias

Amador, J. A., Forns, M. & Martorell, B. (2001a). Síntomas de desatención e hiperactividad-impulsividad: análisis evolutivo y consistencia entre informantes. *Anuario de Psicología, 32*(1), 51-66.

Amador, J. A., Forns, M. & Martorell, B. (2001b). Sensibilidad y especificidad de las valoraciones de

padres y profesores de los síntomas del trastorno por déficit de atención con hiperactividad. *Anuario de Psicología, 32*(4), 65-78.

Amador, J. A., Idiazábal, M. A., Aznar, J. A. & Peró, M. (2003). Estructura factorial de la Escala de Conners para profesores en muestras comunitaria y clínica. *Revista de Psicología General y Aplicada, 56*(2), 173-184.

American Psychiatric Association. (2000). Diagnostic and statistical manual of mental disorders - IV (ed. rev.). Washington, DC: Autor.

Andrich, D. (1978). Rating formulation for ordered response categories. *Psychometrika, 43*(4), 561-573.

Andrich, D. (1988). *Rasch models for measurement* (Series: Quantitative Applications in the Social Sciences, Vol. 07-68). Newbury Park, CA: Sage Publications, Inc.

Ayala, L. J. (2009). *The theory and practice of item response theory*. New York: Guilford Press.

Barkley, R. A. (1990). *Attention-Deficit Hyperactivity Disorder: A handbook for diagnosis and treatment*. New York: Guilford.

Barkley, R. A. (2006). *Attention-Deficit Hyperactivity Disorder: A handbook for diagnosis and treatment* (3a. ed.). New York: Guilford.

Barkley, R. A. & Murphy, K. (2006). *Attention-Deficit Hyperactivity Disorder. A clinical workbook*. New York: Guilford.

Bond, T. G. & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Erlbaum.

Burns, G. L., Walsh, J. A. & Gomez, R. (2003). Convergent and discriminant validity of trait and source effects in ADHD-inattention and hyperactivity-impulsivity measures across a 3-month interval. *Journal of Abnormal Child Psychology, 31*(5), 529-541.

Cardo, E. & Servera, M. (2005). Prevalencia del trastorno de déficit de atención e hiperactividad. *Revista de Neurología, 40*(Supl. 1), S11-S15.

Cohen, J. (1988). *Statistical power for the behavioral sciences* (2a ed.). Hillsdale, NJ: Erlbaum.

Conners, C. K. (1989). *Conners' Rating Scales*. Toronto, Ontario: Multi-Health Systems.

Conners, C. K. (1997). *Conners' Rating Scales: Revised technical manual*. North Towanda, NY: Multi-Health Systems.

Crocker, L. & Algina, J. (2008). *Introduction to classical and modern test theory*. Mason, OH: Cengage Learning.

DuPaul, G. J., Power, T. J., Anastopoulos, A. D. & Reid, R. (1998). *ADHD Rating Scale IV: Checklists, norms, and clinical interpretation*. New York: Guilford.

Embretson, S. E. & Hershberger, S. L. (1999). *The new rules of measurement*. Mahwah, NJ: Erlbaum.

Embretson, S. E. & McCollam, K. M. S. (2000). Psychometric approaches to understanding and measuring intelligence. En R. J. Sternberg (Ed.), *Handbook of intelligence* (pp. 423-444). Cambridge, UK: Cambridge University Press.

Embretson, S. E. & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Farré-Riba, A. & Narbona, J. (1997). Escalas de Conners en la evaluación del trastorno por déficit de atención con hiperactividad: nuevo estudio factorial en niños españoles. *Revista de Neurología, 25*(138), 200-204.

Fidalgo, A. M. (2005). Enfoque de la teoría de respuesta a los ítems. En J. Muñiz, A. M. Fidalgo, E. García-Cueto, R. Martínez & R. Moreno (Eds.), *Análisis de los ítems* (pp. 79-131). Madrid: La Muralla.

Gadow, K. D. & Sprafkin, J. (1997). *Symptom Checklist-4 manual*. Stoony Brook, NY: Checkmate Plus.

Gomez, R. (2007). Testing gender differential item functioning for ordinal and binary scored parent rated ADHD symptoms. *Personality and Individual Differences, 42*(4), 733-742.

Gumpel, T., Wilson, M. & Shalev, R. (1998). An item response theory analysis of the Conners Teacher's Rating Scale. *Journal of Learning Disabilities, 31*(6), 525-532.

Hambleton, R. K., Swaminathan, H. & Rogers, H. J. (1991). *Fundamentals of item response theory*. Beverly Hills, CA: Sage.

LimeSurvey. (2009). LimeSurvey, v. 1.87 [Software de cómputo]. Disponible en http://www. limesurvey. org

Linacre, J. M. (1999). Investigating rating scale category utility. *Journal of Outcome Measurement, 3*(2), 103-122.

Linacre, J. M. (2011). *A user's guide to WINSTEPS, v. 3.73.* Chicago: Winsteps.

Linacre, J. M. & Wright, B. D. (1999). WINSTEPS: Multiple choice, rating scale, and partial credit Rasch analysis [Software de cómputo]. Chicago: MESA Press.

Merrell, K. W. (2003). *Behavioral, social and emotional assessment of children and adolescents* (2a. ed.). Mahwah, NJ: Erlbaum.

Prieto, G. & Delgado, A. R. (1999). Medición cognitiva de las aptitudes. En J. Olea, V. Ponsoda & G. Prieto (Eds.), *Tests informatizados: fundamentos y aplicaciones* (pp. 207-226). Madrid: Pirámide.

Prieto, G. & Delgado, A. R. (2000). Utilidad y representación en la psicometría actual. *Metodología de las Ciencias del Comportamiento, 2*(2), 111-127.

Prieto, G. & Delgado, A. R. (2003). Análisis de un test mediante el modelo de Rasch. *Psicothema, 15*(1), 94-100.

Prieto, G. & Dias, A. (2004). Uso del modelo de Rasch para poner en la misma escala las puntuaciones de distintos tests. *Actualidades en Psicología, 19*(106), 5-23.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago: MESA Press.

Rasch, G. (1977). On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements. En M. Blegvad (Ed.), *The Danish Yearbook of Philosophy* (pp. 58-94). Copenhagen: Munksgaard.

Reid, R., DuPaul, G. J., Power, T. J., Anastopoulos, A., Rogers-Adkinson, D., Noll, M. B., et al. (1998). Assessing culturally different students for attention deficit hyperactivity disorder using behavior rating scales. *Journal of Abnormal Child Psychology, 26*(3), 187-198.

Swanson, J. (2010). *The SNAP-IV Rating Scale.* Disponible en http://www.adhd.net

Wolraich, M., Lambert, E. W., Doffing, M. A., Bickman, L., Simmons, T. & Worley, K. A. (2003). Psychometric properties of the Vanderbilt ADHD diagnostic parent rating scale in a referred population. *Journal of Pediatric Psychology, 28*(8), 559-567.

Wright, B. D. & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement*. Chicago: MESA Press.

Wright, B. D. & Stone, M. H. (1979). *Best test design: Rasch measurement*. Chicago: MESA Press.