Revista Latinoamericana de Psicología (2016) 48, 175–182

Revista Latinoamericana de Psicología

www.elsevier.es/rlp

ORIGINAL

Estudio comparativo del acuerdo y consistencia intercalificadores en el test gestáltico visomotor de Bender 2.ª edición

César Merino Soto[a,*], Gustavo Calderón De la Cruz[a] y Eduardo Manzanares Medina[b]

[a] *Instituto de Investigación de Psicología, Universidad de San Martín de Porres, Lima, Perú*
[b] *Universidad Peruana de Ciencias Aplicadas, Lima, Perú*

Recibido el 30 de julio de 2014; aceptado el 30 de septiembre de 2015
Disponible en Internet el 27 de abril de 2016

PALABRAS CLAVE
Test de Bender;
Confiabilidad;
Acuerdo;
Habilidad visomotora;
Evaluación

Resumen El estudio tiene por objetivo contrastar la efectividad de calificadores sin entrenamiento específico en el test gestáltico visomotor de Bender, 2.ª edición (Bender-II), usando un método para calificar el grado de exactitud de los dibujos reproducidos propio de este instrumento (sistema de calificación global). Algunos estudios previos han demostrado buenos niveles de confiabilidad intercalificador, pero no se verificó el efecto de la falta de entrenamiento específico. En el estudio participaron 75 niños divididos en dos grupos (34 y 41) de edad y cuatro calificadores (dos estudiantes y dos egresados, todos de psicología). Después de aplicar el test individualmente, los calificadores recibieron la instrucción de puntuar los dibujos usando únicamente el manual como guía, sin interactuar entre ellos. Se hicieron comparaciones intragrupo e intergrupos. Aunque los resultados indicaron algunas diferencias moderadas entre los grupos, principalmente se hallaron niveles altos de acuerdo y consistencia; y comparado con los estudios previos, las diferencias generalmente no fueron sustanciales. Se concluye que el puntaje visomotor del Bender-II puede alcanzar buenos niveles de confiabilidad.

KEYWORDS
Bender test;
Reliability;
Agreement;
Visual-motor skill;
Assessment

Comparative study of inter-rater agreement and consistency in the Bender visual motor gestalt test, 2.nd edition

Abstract This study aims to test the effectiveness of raters without specific training in the Global Scoring System, a method used to score the level of accuracy of the drawings reproduced in the Bender gestalt test, 2.nd edition (Bender -II). Some previous studies have shown good reliability in inter-rater levels, but the effect of the lack of specific training was not verified. The participants were 75 children, divided into two groups

* Autor para correspondencia.
Correo electrónico: sikayax@yahoo.com.ar (C. Merino Soto).

http://dx.doi.org/10.1016/j.rlp.2015.09.011

(34 and 41, both between 8 and 11 years old), and four raters (two students and two graduates in psychology). After applying the test individually, the raters were instructed to rate the reproduced designs using the manual as a guide only, with no interaction between them. A within and between-group comparisons analysis was performed. Although the results showed some moderate differences in the group comparisons, high levels of agreement and consistency were mainly found. When compared with previous studies, the differences were generally not substantial. It can be concluded that the Bender-II visual-motor score can achieve good levels of reliability.

El test gestáltico visomotor de Bender –TGB– (Bender, 1938) ha tenido muchos cambios que han afectado su funcionalidad y estructura, esto hace que los trabajos de validez y confiabilidad no sean necesariamente generalizables entre una y otra versión del mismo (Merino, 2011). Dos de las últimas versiones que han aparecido dentro de la literatura científica y práctica profesional son la segunda edición – Bender-II– (Brannigan & Decker, 2003) y la modificación del sistema evolutivo de calificación –Koppitz-II– (Reynolds, 2007). Ambas versiones proponen nuevos sistemas de calificación, alejándose de los sistemas tradicionales que ponían énfasis en el error en las reproducciones (Merino, 2011; Merino, Allen & Decker, 2013). El Bender-II presenta el sistema de calificación global (SCG), propio de este instrumento y aplicable a la evaluación de la copia y recuerdo en el Bender-II. Este método permite representar la continuidad de la habilidad visomotora en los puntajes de los ítems (Merino et al., 2013), mediante el escalamiento de ordinal para calificar cada ítem, y el enfoque global u holístico que debe aplicar el calificador. Esto difiere sustancialmente de las versiones anteriores como el sistema de Koppitz (1984) para niños, o de Lacks (1999) para adultos. Los lectores pueden dirigirse a otras fuentes (e.g., Brannigan, Decker & Madsen, 2004; Merino et al., 2013) para hallar más información relevante sobre estas diferencias. Otros métodos que mantienen este enfoque de continuidad también se han desarrollado para versiones del TGB modificadas localmente (e.g., en Brasil, el sistema de puntuación gradual; Sisto, Noronha & Santos, 2006); sin embargo, aún no son de uso generalizado.

La presente investigación se enfocará en el Bender-II, una herramienta que aparentemente aún es poco utilizada en investigaciones independientes alrededor del mundo. Efectivamente, puede considerarse que en la literatura científica actual aún hay pocos estudios sobre generalización de la validez y confiabilidad del Bender-II. Los principales estudios de las propiedades psicométricas del Bender-II se han realizado en su manual (Brannigan & Decker, 2003), donde se encuentra un fuerte respaldo del Bender-II para la diferenciación de grupos con discapacidades cognitivas, su concurrencia con otras medidas de visomotricidad y sus correlatos con el rendimiento académico. En las investigaciones posteriores realizadas con la muestra de estandarización americana (Decker, 2007; Decker, Allen & Choca, 2006; Decker, Englund, Carboni & Brooks, 2011) y en las investigaciones independientes en muestras hispanas (Merino, 2012; Merino & Allen, 2012, 2013), anglosajonas (Volker et al., 2010), y en el Medio Oriente (Uluç, Gülüm & Çalişir, 2012), se puede hallar un continuo respaldo favorable a la validez y confiabilidad del Bender-II.

En la evaluación de la confiabilidad intercalificadores, un aspecto crucial es la estimación de la confiabilidad y acuerdo intercalificadores, pues la diferente interpretación que cada evaluador hace del sistema de calificación añade una fuente de variación de error a los puntajes (Cone, 1999; Feldt & Brennan, 1989). Aunque los estudios que evaluaron la consistencia y/o el acuerdo intercalificador del Bender-II reportaron elevados coeficientes (correlaciones lineales e intraclase) para el puntaje total (p. ej. Merino, 2012; Merino & Allen, 2013; Volker et al., 2010) y niveles aceptables para la mayoría de los ítems (Merino, 2012), estos estudios generalmente se caracterizaron por tener calificadores entrenados previamente en el Bender-II. Esta preparación asegura que el diseño del estudio de confiabilidad homogeneice la experiencia en el material (en esta situación, el SCG), pero puede no representar una condición natural en el uso del Bender-II durante la práctica profesional.

En otras palabras, cuando los profesionales o investigadores adquieren un nuevo instrumento, parece más probable que se autoinstruyan en el uso de los materiales y procedimientos de aplicación y calificación. Esta instrucción autodidacta representa la suma de la experiencia propia y los entrenamientos previos en otros instrumentos pueden servir para transferir las habilidades adquiridas hacia el nuevo instrumento. La interacción entre la experiencia, el juicio, el apego a estándares de uso de pruebas (por ejemplo, International Test Commission, 2000), y quizás el ensayo y error juicioso, pueden conducir a que el usuario maneje satisfactoriamente el instrumento. Cuando esto ocurre, uno puede preguntarse si se pueden obtener puntajes de buena calidad, es decir, que contengan poco error o sesgo. Es posible que la edición del manual no sea suficiente para asegurar un aprendizaje exitoso del método holístico de calificación.

En el manual, Brannigan y Decker (2003) afirmaron que el SCG requiere poco entrenamiento para lograr alta consistencia, y que sería suficiente el aprendizaje de las instrucciones del manual, específicamente como práctica autodidacta. Como evidencia de ello, los autores solicitaron que un

evaluador calificara 66 protocolos extraídos aleatoriamente de la muestra de estandarización americana, y sus puntajes en copia y recuerdo fueron comparados con un experto en el uso del Bender- II. Se obtuvieron coeficientes de correlación iguales a 0.85 y 0.92 (respectivamente), lo que respaldó la posición de los autores. Sin embargo, esta estimación solo respaldó la consistencia entre los calificadores pero no el acuerdo entre los mismos, pues la correlación lineal solo es sensible a los cambios en la monoticidad de las respuestas pero no al cambio efectivo o exacto en las puntuaciones. Adicionalmente, un documento de trabajo publicado posteriormente al manual (Brannigan et al., 2004) extendió la práctica de calificación del SCG con un ejemplo detallado, que incluía explicaciones comparativas. Se podría argumentar que esta guía adicional sirvió para complementar la necesidad de obtener mayor información publicada en el manual un año antes, y proporcionar una mayor práctica y comprensión de este nuevo método de calificación. Finalmente, los estudios previos sobre la confiabilidad de la calificación en muestras independientes (Merino, 2012; Merino & Allen, 2013; Volker et al., 2010) obtuvieron elevados coeficientes de consistencia, pero los calificadores recibieron sesiones de entrenamiento; por lo tanto, estos resultados son experimentalmente diferentes.

El presente estudio tiene por finalidad evaluar la calificación del Bender-II mediante calificadores sin entrenamiento en el instrumento y su sistema de calificación. El estudio diferenciará la evaluación de la consistencia y del acuerdo, aspectos conceptualmente distintos (Cone, 1999; Esquivel et al., 2006; Von Eye & Mun, 2005), y que tiene relevancia psicométrica y práctica para el profesional e investigador. La consistencia y el acuerdo entre calificadores representan dos formas de describir la precisión de los puntajes, pero se refieren a dos modelos diferentes que pueden ser confundidos, y no necesariamente coexistir en la misma magnitud (Cone, 1999; Liao, Hunt & Chen, 2010). Mientras la consistencia entre los calificadores apunta hacia el ordenamiento diferenciado de los sujetos (o estímulos), el acuerdo se refiere a la exactitud o concordancia de los puntajes entre los calificadores (Cone, 1999; Liao et al., 2010), y puede ser recomendado que ambos índices se reporten conjuntamente (Fleenor, Fleenor & Grossnickle, 1996) para obtener una clara figura de la similaridad entre observadores. Aunque esta diferenciación no fue reconocida en los estudios originales del Bender-II (Brannigan & Decker, 2003), algunos reportes independientes sobre la precisión de sus puntajes (Merino, 2012; Merino & Allen, 2013) han reportado índices diferenciados de consistencia y acuerdo, pero en un contexto diferente al del presente estudio.

Método

Participantes

Los participantes fueron los niños a quienes se les administró el Bender-II y los calificadores. Lo niños provinieron independientemente de dos instituciones educativas estatales de educación básica regular, ubicados en Lima Metropolitana (Perú); fueron 34 niños en el grupo 1 (15 mujeres) y 41 niños (21 mujeres) en el grupo 2. El rango de edad de ambos grupos estuvo entre 9 y 11 años. Este rango pertenece al segundo nivel de edad de aplicación de las pruebas aplicadas en el presente estudio (es decir, entre 8 años a más; ver sección Instrumentos). Las instituciones de donde provienen los niños fueron elegidas por su disponibilidad para participar en el estudio, y de acuerdo con el juicio de los autores del presente estudio, no parecieron representar instituciones con características especiales de gestión o muy diferenciadas del resto de instituciones educativas en Lima. Para evaluar a los participantes se contó con la autorización del director del centro educativo y de los padres de familia, a quienes se les informó que la evaluación era con fines de investigación.

Por otra parte, los calificadores fueron de la carrera de psicología, pertenecientes a una universidad privada de Lima Metropolitana. Fueron cuatro sujetos distribuidos en el grupo 1 y grupo 2 anteriores, que se encargaron de administrar el Bender-II y calificarlos posteriormente en sus grupos, y dentro del marco del presente estudio. En el grupo 1 los calificadores fueron dos recientes egresados de psicología, que laboraban en instituciones educativas y con experiencia en el uso de pruebas visomotoras anteriores (y exceptuando también) al Bender-II. En el grupo 2 se incluyeron dos estudiantes de Psicología de quinto ciclo, con alguna experiencia en una versión antigua del TGB (e.g., Koppitz, 1984). Los cuatro calificadores tuvieron un rendimiento promedio o mayor que el promedio durante sus estudios.

Instrumento

Test gestáltico visomotor de Bender, segunda edición –Bender-II–, (Brannigan & Decker, 2003). Es la nueva versión del BGT que consta de 16 diseños, aplicable de 4 a 85 años de edad; un conjunto diferente de diseños se administra para el nivel de edad desde 4 a 8 años, y otros diseños para el rango de edad de 8 a 85 años. Evalúa la habilidad visomotora, y mediante pruebas complementarias examina otras habilidades cognitivas relacionadas con la visomotricidad. Por lo tanto, su aplicación consiste en una fase copia (visomotricidad), de recuerdo (memoria visual constructiva: el evaluado vuelva a reproducir los diseños que pueda recordar), motora (control motor: trazos continuos uniendo los puntos extremos) y perceptual visual (discriminación visual: identificación de figuras). La calificación de las reproducciones en la copia y el recuerdo se realiza con el SCG, el cual consiste en calificar cada uno de los diseños desde una escala de 0 (*ausencia de forma en el dibujo*) hasta 4 (*dibujo casi perfecto*). Este método enfatiza la calidad global del diseño y forma parte del propio instrumento. En el manual (Brannigan & Decker, 2003), la consistencia interna varía entre 0.86 a 0.95; y la correlation test-retest entre 0.80 y 0.88, para la fase de copia; y de 0.80 a 0.86, para la fase de recuerdo. Los estudios iniciales de validación en muestras hispanas (Merino, 2012; Merino & Allen, 2012, 2013) han reportado satisfactorias propiedades psicométricas en niños. La información detallada y comparativa del Bender-II respecto a sus características puede hallarse en Merino et al. (2013) o Brannigan et al. (2004).

Procedimiento

Después de recibir la autorización de la institución educativa y de los padres de familia de los niños, se inició la

aplicación del Bender-II. Los niños que aprobaron participar voluntariamente, fueron citados para la sesión individual de evaluación con el Bender-II. Se usó la traducción previa de las instrucciones y procedimiento de uso (Merino, 2012), las instrucciones de aplicación descritas en el manual (Brannigan & Decker, 2003) se mantuvieron intactas. La aplicación del Bender-II se hizo de manera individual, en una sola sesión de evaluación, y dentro de un aula de clases vacía. Los examinadores fueron los mismos que participaron en la calificación de los diseños reproducidos. El procedimiento para la calificación fue entregar a los calificadores los protocolos con la instrucción de que revisaran el manual para guiarse en la calificación de los dibujos, y que las posibles dudas deberían resolverlas en el mismo documento, sin interactuar entre ellos ni solicitar asesoría del investigador principal del presente estudio. No se dieron más instrucciones respecto a la calificación y se enfatizó que evitaran alguna interacción durante el proceso de calificación. La calificación de los protocolos les tomó aproximadamente dos semanas y media a cada calificador.

Análisis de datos

Para el análisis estadístico, primero se hicieron análisis preliminares y se obtuvieron la consistencia interna (Cronbach, 1951) y sus intervalos de confianza (Dominguez & Merino, 2015); también se compararon los puntajes promedio de los calificadores en los grupos 1 y 2, mediante la prueba *t* de comparación de medias dependientes, y la propuesta de Dunlap, Cortina, Vaslow y Burke (1996) para estimar la magnitud de las diferencias. Esta toma en cuenta la correlación entre las medias en su formulación:

$$d = t_c\sqrt{\frac{2\,(1-r)}{n}} \tag{1}$$

Para los análisis principales, se aplicaron varios procedimientos para estimar y examinar las diferencias en la consistencia y acuerdo los puntajes (copia y recuerdo) del Bender-II. En cada análisis se hizo énfasis en la significación práctica para valorar con más información las diferencias (Ledesma, Macbeth & Cortada, 2008). Se estimó la consistencia mediante correlaciones Pearson, y el acuerdo mediante coeficientes de correlación intraclase (ICC); este último se hizo con el método de dos factores aleatorios (McGraw & Wong, 1996). Luego, se evaluaron las diferencias en la consistencia (correlaciones Pearson) y en el acuerdo (ICC), ambos en un marco de comparación intragrupo (diferencia entre el acuerdo del puntaje copia y recuerdo en un mismo grupo) e intergrupo (diferencia entre los grupos 1 y 2 respecto puntajes de copia y recuerdo). En el análisis intergrupo para la diferencia entre correlaciones, se usó la prueba *Z* para dos grupos independientes (Chen & Popovich, 2002); y en el análisis intragrupos, se aplicó una prueba de contraste para dos grupos dependientes sin elementos comunes, Z_{PF} (Raghunathan, Rosenthal & Rubin, 1996). En ambas comparaciones intergrupo e intragrupo, se usó el estimador *q* para ver la magnitud del efecto (Cohen, 1992), tomando en cuenta los siguientes niveles: trivial ($< \pm 0.20$), baja ($\geq \pm 2.10$), moderada ($\geq \pm 0.50$), alta ($\geq \pm 0.80$).

Para examinar las diferencias en el acuerdo (coeficientes ICC), se usó la prueba de contraste para muestras independientes (comparación intergrupo; Alsawalmeh & Feldt, 1992) y dependientes (comparación intragrupo; Alsawalmeh & Feldt, 1994), ejecutada por un programa *ad hoc* (Merino, 2013). La magnitud del efecto de estas comparaciones se realizó transformando los coeficientes ICC a valores estandarizados, mediante la transformación a z de Fisher propuesta por Konishi (1985), un método que aproxima bien los coeficientes ICC a valores normalizados (Donner & Zou, 2002):

$$Zm = \left[\sqrt{\frac{k-1}{2k}}\,in\left(\frac{1+(k-1)\,ICC}{1-ICC}\right)\right] + \frac{7-5k}{N\sqrt{18k\,(k-1)}} \tag{2}$$

En la fórmula, *k* es el número de calificadores en el cálculo de ICC, *N* es el tamaño de la muestra. Como nota procedimental, el contraste entre ICC intragrupo requiere la correlación entre los puntajes; por lo tanto, esta correlación conjunta entre copia y memoria dentro de cada grupo 1 y 2, fue estimada promediando las correlaciones entre ambos puntajes en cada evaluador. De este modo, en el grupo 1 la correlación copia-recuerdo en el calificador E fue 0.477 y en el calificador G fue 0.538; entonces, el promedio fue 0.50. En el grupo 2, la correlación copia-recuerdo en el calificador F fue 0.246 y en el calificador D fue 0.354; entonces, el promedio fue 0.30. El contraste entre ICC se hizo también con el estudio de Merino (2012), quien obtuvo coeficientes ICC en una muestra entrenada de calificadores. Los niveles de acuerdo elegidos para describir los resultados cualitativamente, fueron los elaborados por Cicchetti (1994), que parecen ser los más popularizados en la literatura actual (Hallgren, 2012). Estos niveles son: $<$ 0.40: pobre; $\geq$ 0.40: aceptable; $\geq$ 0.60: bueno; $\geq$ 0.75: excelente.

Resultados

Análisis preliminar

Los resultados descriptivos aparecen en la tabla 1. Se observa que en el grupo 1, las diferencias entre los calificadores en copia ($t[33] = 6.87$, $r = 0.90$) y recuerdo ($t[33] = 3.34$, $r = 0.95$) fueron estadísticamente significativas, pero de baja ($d = 0.526$) y trivial ($d = 0.181$) magnitud. Por otro lado, en el grupo 2 las diferencias en el puntaje copia fueron estadísticamente significativas ($t[40] = 2.57$, $r = 0.92$) pero trivial en magnitud ($d = 0.160$), mientras que en recuerdo no se detectó significación estadística ($t[40] = 0.36$, $r = 0.87$, $d = 0.02$). Los resultados descriptivos entre el grupo 1 y 2 no son comparables pues la muestra de protocolos en ambos fue diferente para cada uno. Sobre la consistencia interna por el método α (Cronbach, 1951), se hallaron coeficientes superiores a 0.70, aunque más bajos en el grupo 2 comparado con el grupo 1 (tabla 1). Los intervalos de confianza (90%) fueron estimados con el método Bonnett (2002).

Consistencia intergrupo e intragrupo

Excepto para el puntaje de recuerdo en el grupo 2, en ambos grupos de calificadores la consistencia de los puntajes fue elevada (correlaciones Pearson ε 0.90) para los puntajes

Tabla 1 Estadísticos descriptivos para los puntajes del Bender–II

	M	DE	α [IC 90%]	Estadísticos de distribución				
				Simetría		Curtosis		Sw
				As.[a]	Z	Cu.[b]	Z	
Grupo 1 (*n* = 34)								
Copia								
E	35.08	5.43	0.78 [0.66, 0.86]	-0.27	-0.69	-0.50	-0.645	0.96
G	37.88	4.73	0.75 [0.61, 0.84]	-0.19	-0.47	-0.85	-1.080	0.96
Recuerdo								
E	15.38	6.14	-	0.26	0.64	1.92	2.440	0.96
G	16.58	6.92	-	0.63	1.56	2.38	3.028	0.95
Grupo 2 (n = 41)								
Copia								
F	38.73	4.08	0.71 [0.57, 0.80]	-0.40	-0.55	-0.32	-0.875	0.96
D	38.10	4.17	0.71 [0.57, 0.80]	-0.25	-0.35	-0.58	-1.572	0.97
Recuerdo								
F	17.00	7.00	-	0.133	0.184	-0.50	-1.379	0.98
D	17.20	6.77	-	0.047	0.065	-0.20	-0.556	0.98

Grupo 1: calificadores-examinadores egresados; Grupo 2: calificadores-examinadores estudiantes. E, G, D y F: letras para diferenciar a los calificadores.
[a] coeficiente de asimetría.
[b] coeficiente de curtosis.

copia y recuerdo (tabla 2). En el contraste intergrupos, únicamente se detectaron diferencias en el puntaje recuerdo ($p < .05$), la misma que puede considerarse pequeña; en cambio, en el puntaje copia, el tamaño de la diferencia intergrupos fue trivial ($p > .05$). Al examinar el contraste intragrupos (entre copia y recuerdo), en el grupo 1 se detectaron diferencias moderadas ($p > .05$), y en el grupo 2 el tamaño de la diferencia fue pequeña; en ambos, las diferencias no fueron estadísticamente significativas ($p > .10$).

Comparando la consistencia promedio de copia (0.91) con la obtenida (correlación promedio = 0.906) del estudio en Merino (2012), no se halló una diferencia estadísticamente significativa ($Z = -0.108$, $p > .05$, $q = 0.02$). De manera similar, su correlación promedio en recuerdo (0.90; Merino, 2012) con el promedio obtenido del presente estudio (0.91) para este mismo puntaje, tampoco fue sustancialmente diferente ($Z = -0.263$, $p > .50$, $q = 0.05$). Finalmente, la consistencia entre los calificadores no se pudo comparar con los resultados de Merino (2012), pues ahí se usó una modificación del coeficiente ICC para consistencia, no correlaciones Pearson.

Acuerdo intergrupo e intragrupo

En la tabla 3, el contraste intergrupos mostró diferencias estadísticas únicamente en el puntaje recuerdo, donde la discrepancia puede considerarse grande; mientras que en el puntaje copia, la diferencia puede considerarse moderada aunque estadísticamente no significativa. En el contraste intragrupo, se halló que en el grupo 1 hay diferencias más allá del error de muestreo, y de magnitud grande; en cambio, en el grupo 2, la discrepancia no fue estadísticamente significativa y fue de baja magnitud. El nivel de acuerdo es excelente en todos los coeficientes de la tabla 3.

Al comparar los coeficientes ICC obtenidos con los del estudio de Merino (2012), que fue 0.85 en copia y recuerdo, para el puntaje copia se hallaron diferencias más allá del error de muestreo en el grupo 1 ($F[42,18] = 2.14$, $p < .05$, $q = 0.80$), pero no en el grupo 2 ($F[43,21] = 1.15$, $p > .05$, $q = 0.15$). En el puntaje recuerdo, no hubo diferencias de importancia estadística y práctica en el grupo 1 ($F[37,21] = 1.44$, $p > .05$, $q = 0.38$) y en el grupo 2 ($F[44,20] = 1.66$, $p > .05$, $q = 0.54$).

Tabla 2 Consistencia (correlación Pearson) entre e intragrupos para el Bender-II

	Grupo 1 (*n* = 34)	Grupo 2 (*n* = 41)	Diferencias entre grupo	
Puntajes				
Copia [IC 95%]	0.90 [0.80, 0.94]	0.92 [0.85, 0.95]	$Z = -0.48$	$q = 0.11$
Recuerdo [IC 95%]	0.95 [0.90, 0.97]	0.87 [0.76, 0.92]	$Z = 2.06^{*}$	$q = 0.49$
Diferencias intragrupo	$Z_{pf} = -1.50$ $q = 0.35$	$Z_{pf} = 1.11$ $q = 0.25$		

q: estimación de magnitud del efecto; *Z*: prueba de diferencia de correlaciones independientes; Z_{pf}: prueba de diferencia de correlaciones dependientes.
[*] $p < .05$.

Tabla 3 Acuerdo (coeficiente ICC) entre e intragrupos para el Bender-II

	Grupo 1 ($n = 34$)	Grupo 2 ($n = 41$)	Diferencias intergrupo
Puntajes			
Copia [IC 95%]	0.93 [0.825, 0.971]	0.87 [0.77, 0.93]	$F(39, 20) = 1.85$ $q = 0.64$
Recuerdo [IC 95%]	0.77 [0.078, 0.927]	0.91 [0.83, 0.95]	$F(35, 22) = 2.55^{*}$ $q = 1.01$
Diferencias intragrupo	$F(38, 22) = 3.28^{**}$ $q = 1.27$	$F(42, 23) = 1.44$ $q = 0.38$	

q: estimación de magnitud del efecto.
* $p < .05$.
** $p < .01$.

Discusión

Los resultados del presente estudio demuestran que se pueden obtener buenos niveles de consistencia entre los puntajes, que en general están alrededor de 0.90. Esta magnitud generalmente se sugiere para que un instrumento psicológico pueda utilizarse en la descripción individual, y tomar decisiones en conjunto de otras mediciones relevantes (Nunnally & Bernstein, 1995). Se detectaron diferencias triviales en la consistencia del puntaje copia que llevan a esperar que estas tengan poco valor práctico, de tal modo que la posición de un examinado puede ser casi exactamente igual en los puntajes provenientes de dos examinadores sin experiencia en el uso del Bender-II. Por otro lado, en el puntaje de recuerdo se observaron diferencias pequeñas, sugiriendo que las discrepancias sí podrían hallarse en este puntaje usando los criterios del SCG. Estas discrepancias no triviales (pero bajas) deben valorarse en el contexto del uso del Bender-II, pues la varianza de error tiene consecuencias más serias cuando se requiere tomar decisiones individuales sobre los niños evaluados, en contraste con el uso del Bender-II en la investigación básica o aplicada (International Test Commission ITC, 2000).

Esta pequeña inconsistencia representa la falta de linealidad entre las calificaciones, especialmente en alguna parte de la distribución de puntajes de recuerdo, pero en este estudio no se verificó la ubicación de esta discrepancia. Por otro lado, desde el enfoque ANOVA (mediante la ICC), el acuerdo intercalificadores en evaluadores sin entrenamiento puede alcanzar niveles que en la literatura metodológica son considerados como altos (por ejemplo, ε 0.75; Cicchetti, 1994); para el puntaje copia, estas magnitudes en el presente estudio superaron el valor 0.85 en la muestra y 0.75 en la población. Para el puntaje de recuerdo, el grado de acuerdo entre los dos grupos tuvo fuertes discrepancias, así como en la diferencia intragrupo para uno de los grupos.

Se puede conjeturar que las discrepancias en el puntaje de recuerdo pueden ocurrir en las zonas bajas de puntuación, donde parece ser más difícil la aplicación del SCG. Efectivamente, el dibujo reproducido con baja puntuación (p. ej. 1 o 2) puede representar un diseño bien recordado pero pobremente diseñado, problema que confundiría al usuario ya que el criterio del SCG apunta hacia la calidad de la reproducción, sin tomar en cuenta el contexto de memorización. En esta situación, el usuario sin entrenamiento podría hallar que el criterio que elige es razonable, mientras que otro usuario podría elegir otro. Para explorar esto, se requieren estudios que planteen esta hipótesis, y dar el paso adicional para identificar la ubicación del desacuerdo. Procedimientos como el método gráfico Bland-Altman (Bland & Altman, 1999; Bland & Altman, 1986) pueden ayudar a diferenciar el tipo de sesgo ocurrido, además de detectar si el sesgo es proporcional a la magnitud de los puntajes. Estos métodos proporcionan información que serviría para mejorar el diseño de los estudios de confiabilidad y la enseñanza de pruebas psicológicas proclives al error de medición entre calificadores/observadores. Aun con la discrepancia ocurrida en el puntaje de recuerdo, su magnitud no es baja sino sobre 0.75, y tomando en cuenta la falta de conocimiento previo sobre el Bender-II, no parece ser un inconveniente para afirmar que el nivel de acuerdo es aceptable para propósitos prácticos. Algunos estudios han reportado correlaciones intercalificadores de 0.90 (Friedman, Fuerth & Forsythe, 1980).

Un aporte adicional del presente estudio es la propuesta de un indicador de magnitud del efecto basado en el trabajo de Donner y Zou (2002) y Konishi (1985), que produce una estadístico estandarizado basado en la transformación z de los coeficientes ICC. Esta transformación tiene buenas propiedades estadísticas y es adaptable a varias situaciones (Sánchez-Bruno & Borges del Rosal, 2005), por ejemplo como la del presente estudio. La racionalidad del este método descansa en el proceso básico de estandarización del estadístico de diferencias de ICC transformados a unidades z Konishi (1985) y la comparación de ellos propuesto por Cohen (1992). Sin embargo, los límites para establecer cualitativamente la magnitud de las diferencias (trivial, bajo, moderado y alto) están abiertos a debate.

Respecto a las limitaciones en el presente estudio, solo dos calificadores participaron en cada grupo, lo cual es una situación característica que podría representar la práctica profesional de psicólogos o equipos de trabajo en una institución educativa. Esto condiciona a que el psicólogo sea el único que administre y califique los instrumentos, y que posiblemente cuente con otro colega o asistente que participe del proceso. Por lo tanto, la limitación de contar únicamente con dos calificadores en cada grupo tendría la ventaja de ser ecológicamente válido.

Otra posible limitación es el tamaño de la muestra de dibujos en cada grupo, que es comparativamente pequeño frente a los estudios de pruebas visomotoras comerciales; por ejemplo, en el estudio de validación de la prueba Beery-Buktenica del desarrollo de la integracion visomotora; Beery, 2000), se reportan muestras de alrededor de 130. Sin embargo, en el Bender-II, el tamaño muestral varía entre 30 y 100 (Brannigan & Decker, 2003; Merino & Allen, 2013; Uluç et al., 2012; Volker et al., 2010).

Aunque nuestro estudio obtiene un excelente poder estadístico en el nivel de correlación mínima esperada para el Bender-II (0.80, prueba de una cola, *n* promedio = 37, correlación nula de 0.0); frente a un razonable *r* nulo de 0.60, el poder alcanzado es 77% para un poder estadístico de 80%, el mínimo *n* es 40, un valor cercano a una de las muestras de estudio. En conjunto, las limitaciones del estudio plantean la necesidad de un estudio de replicación completa o parcial, en el que se consideren calificadores con otras características (experiencia de trabajo o docente, por ejemplo) y quizás un mayor tamaño muestral (de calificadores y protocolos).

Se concluye que, en personas sin entrenamiento en el Bender-II se puede conseguir altos coeficientes de consistencia y acuerdo entre calificadores usando el SCG del Bender-II para el puntaje copia, y ligeramente bajo para el puntaje de recuerdo. Sin embargo, si se requiere asegurar más precisión en la obtención de los puntajes, entonces una o dos sesiones de entrenamiento son necesarias para maximizar la varianza verdadera alrededor del puntaje obtenido, lo que es particularmente más importante para obtener el puntaje de la subprueba de recuerdo, que parece ser más complejo al aplicar el SCG.

Conflicto de intereses

Los autores declaran no tener ningún conflicto de intereses.

Referencias

Alsawalmeh, Y. M. & Feldt, L. S. (1992). Test of the hypothesis that the intraclass coefficient is the same for two measurement procedures. *Applied Psychological Measurement*, *16*, 195–205. http://dx.doi.org/10.1177/014662169201600208

Alsawalmeh, Y. M. & Feldt, L. S. (1994). Testing the equality of two related intraclass reliability coefficients. *Applied Psychological Measurement*, *18*(2), 183–190. http://dx.doi.org/10.1177/014662169401800207

Bender, L. A. (1938). *A visual motor gestalt test and its clinical use*. New York. NY: American Orthopsychiatric Association.

Beery, K. (2000). *Prueba Beery-Buktenica del desarrollo de la integración visomotriz VMI*. México, DF: El Manual Moderno.

Bland, J. M. & Altman, D. G. (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research*, *8*, 135–160. http://dx.doi.org/10.1177/096228029900800204

Bland, J. M. & Altman, D. G. (1986). Statistical method for assessing agreement between two methods of clinical measurement. *The Lancet*, (i), 307–310. http://dx.doi.org/10.1016/S0140-6736(86)90837-8

Bonett, D. G. (2002). Sample size requirements for testing and estimating coefficient alpha. *Journal of Educational and Behavioral Statistics*, *27*, 335–340. http://dx.doi.org/10.3102/10769986027004335

Brannigan, G. G. & Decker, S. L. (2003). *Bender visual-motor gestalt test* (2.ª. ed.). Ithaca, IL: Riverside Publishing.

Brannigan, G. G., Decker, S. L. & Madsen, D. H. (2004). *Innovative features of the Bender-Gestalt II and expanded guidelines for the use of the global scoring system* (2.nd ed.). Itasca, IL: Riverside Publishing (Bender visual-motor gestalt test, assessment service bulletin N.º 1).

Chen, P. Y. & Popovich, P. M. (2002). *Correlation: Parametric and nonparametric measures*. Thousand Oaks. CA: Sage.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*, 284–290. http://dx.doi.org/10.1037/1040-3590.6.4.284

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. http://dx.doi.org/10.1037/0033-2909.112.1.155

Cone, J. D. (1999). Observational assessment: Measure development and research issues. En P. C. Kendall, J. N. Burcher, & G. N. Holmbeck (Eds.), *Handbook of research methods in clinical psychology* (2.nd ed., pp. 183–223). New York, NY: John Wiley & Sons.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of test. *Psychometrika*, *16*, 297–334. http://dx.doi.org/10.1007/BF02310555

Decker, S. L. (2007). Measuring growth and decline in visual-motor processes using the Bender Gestalt II. *Psychoeducational Assessment*, *26*(1), 3–15. http://dx.doi.org/10.1177/0734282907300685

Decker, S. L., Allen, R. & Choca, J. P. (2006). Construct validity of the Bender-Gestalt II: Comparison with Wechsler Intelligence Scale for Children-III. *Perceptual and Motor Skills*, *102*(1), 133–141. http://dx.doi.org/10.2466/pms.102.1. 133-141

Decker, S. L., Englund, J. A., Carboni, J. A. & Brooks, J. H. (2011). Cognitive and developmental influences in visual-motor integration skills in young children. *Psychological Assessment*, *23*(4), 1010–1016. http://dx.doi.org/10.1037/a0024079

Dominguez, S. & Merino, C. (2015). ¿Por qué es importante reportar los intervalos de confianza del coeficiente alfa de Cronbach? *Revista Latinoamericana de Ciencias Sociales, Niñez y Juventud*, *13*(2), 1326–1328.

Donner, A. & Zou, G. (2002). Testing the equality of dependent intraclass correlation coefficients. *Journal of the Royal Statistical Society (Series D)*, *51*(3), 367–379. http://dx.doi.org/10.1111/1467-9884.00324

Dunlap, W. P., Cortina, J. M., Vaslow, J. B. & Dunlap, M. J. (1996). Meta-Analysis of experiments with matched groups or repeated measures designs. *Psychological Methods*, *1*(2), 170–177. http://dx.doi.org/10.1037/1082-989x.1.2.170

Esquivel, C., Velasco, V., Martínez, E., Barbachano, E., Gonzáles, G. & Castillo, C. (2006). Coeficiente de correlación intraclase vs.correlación de Pearson de la glucemia capilar por reflectometría y glucemia plasmática. *Medicina Interna de México*, *22*(3), 165–171.

Feldt, L. S. & Brennan, R. (1989). Reliability. En R. L. Linn (Ed.), *Educational measurement* (3.rd ed., pp. 105–146). New York, NY: Macmillan.

Fleenor, J. W., Fleenor, J. B. & Grossnickle, W. F. (1996). Interrater reliability and agreement of performance ratings: A methodological comparison. *Journal of Business and Psychology*, *10*(3), 367-.80. doi: 10.1007/BF02249609.

Friedman, R., Fuerth, J. H. & Forsythe, A. B. (1980). A brief screening battery for predicting school achievement at ages seven and nine years. *Psychology in Schools*, *17*(3), 340–346, 10.1002/1520-6807(198007)17:3<340::AID-PITS2310170310>3.0.CO;2-7.

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, *8*(1), 23–34.

International Test Commission (ITC; 2000). Guidelines on test use: Spanish version. Translation authorized by the Colegio Oficial de Psicólogos. ITC: Autor.

Konishi, S. (1985). Normalizing and variance stabilizing transformation for intraclass correlations. *Annals of Institute of Statistical Mathematical*, *37*(A), 87–94. http://dx.doi.org/10.1007/BF02481082

Koppitz, E. (1984). *El test gestáltico vismotor para niños* (10.ª ed.). Buenos Aires, Argentina: Guadalupe.

Lacks, P. (1999). *Bender Gestalt screening for brain dysfunction* (2.ª ed.). New York, NY: John Wiley y Sons.

Ledesma, R., Macbeth, G. & Cortada, N. (2008). Tamaño del efecto: revisión teórica y aplicaciones con el sistema estadístico Vista. *Revista Latinoamericana de Psicología*, *40*(3), 425–439.

Liao, S. C., Hunt, E. A. & Chen, W. (2010). Comparison between inter-rater reliability and inter-rater agreement in performance assessment. *Annals Academy of Medicine*, *39*(8), 613–618.

McGraw, K. O. & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*(1), 30–46. http://dx.doi.org/10.1037/1082-989X.1.4.390

Merino, C. (2011). *El Test gestáltico de Bender: Evolución y nuevas versiones. Trabajo presentado en el III Congreso Internacional de Psicología*. Lima, Perú: Universidad Autónoma del Perú.

Merino, C. (2012). Fiabilidad en el test gestáltico de Bender-II, en una muestra independiente de calificadores. *Revista de Investigación Educativa*, *30*(1), 223–234.

(2013). ICC compare: Un programa MS Excel para probar la igualdad de coeficientes de correlación intraclase. Manuscrito inédito.

Merino, C., & Allen, R. A. (2012). A factor-analytic study for the Bender gestalt test, 2.nd edition: Internal structure and measurement model. Cartel presentado en el 30 International Congress of Psychology, Cape Town, South Africa.

Merino, C. & Allen, R. A. (2013). Confiabilidad intercalificadores y validez de constructo del Test Gestáltico Visomotor de Bender (segunda versión). *Interdisciplinaria*, *30*(2), 253–264.

Merino, C., Allen, R. A., & Decker, S. L. (2013). Test gestáltico visomotor de bender – 2ª versión. *Liberabit, 12(2)*, 275-278.

Nunnally, J. & Bernstein, I. (1995). *Teoría psicométrica* (3.ª ed.). México, D.F: McGraw-Hill.

Raghunathan, T. E., Rosenthal, R. & Rubin, D. B. (1996). Comparing correlated but nonoverlapping correlations. *Psychological Methods*, *1*, 178–183. http://dx.doi.org/10.1037/1082-989X.1.2.178

Reynolds, C. (2007). *Koppitz developmental scoring system for the Bender gestalt test (KOPPITZ-2)*. Austin, TX: Pro-Ed.

Sánchez-Bruno, A. & Borges del Rosal, A. (2005). Transformación z de Fisher para la determinación de intervalos de confianza del coeficiente de correlación de Pearson. *Psichotema*, *17*(1), 148–153.

Sisto, F., Noronha, A. & Santos, A. (2006). *Teste gestaltico visomotor de Bender–Sistema de Pontuaçao Gradual (B-SPG)*. Sao Paulo. Brasil: Vetor.

Uluç, S., Gülüm, I. V. & Çalişir, M. (2012). Inter-rater reliability of the Bender visual motor gestalt coordination test (2a ed.) for global, koppitz and recall scoring systems [Hindu]. *Klinik Psikiyatri*, *15*, 71–79.

Volker, M. A., Lopata, C., Vujnovic, R. K., Smerbeck, A. M., Toomery, J. A., Rodgers, J. D., et al. (2010). Comparison of the Bender Gestalt-II and VMI-V in samples of typical children and children with high-functioning autism spectrum disorders. *Journal of Psychoeducational Assessment*, *28*(3), 187–200. http://dx.doi.org/10.1177/0734282909348216

Von Eye, A. & Mun, E. Y. (2005). *Analyzing rater agreement: Manifest variable methods*. Mahwah. NJ: Lawrence Erlbaum.