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

FIABILIDAD EN EL TEST GESTÁLTICO DE BENDER – II, EN UNA MUESTRA INDEPENDIENTE DE CALIFICADORES

César A. Merino Soto
Instituto de Investigación en Psicología, Universidad de San Martín de Porres

RESUMEN

El presente trabajo examina exploratoriamente el acuerdo entre calificadores para la nueva versión del Test Gestáltico Visomotor de Bender – 2da versión (Bender-II). Esta nueva versión tiene importantes cambios en el modo de calificación y administración y aún no se han publicado investigaciones psicométricas en habla hispana. Este estudio es uno de los primeros en que analizará el error de medición de los puntajes del Bender-II en una muestra independiente. La muestra fue de niños de primer y segundo grado de primaria de un colegio privado, y 4 calificadores estudiantes de una universidad privada. Se analizó el grado de concordancia intercalificadores (correlación intraclase y Pearson), así como la consistencia interna (α de Cronbach). Se obtuvieron elevados coeficientes, tanto para la consistencia interna (α de Cronbach > 0.80), y para el acuerdo intercalicadores en la muestra total (ICC > 0.85) y en las comparaciones pareadas (ICC > 0.72) de los puntajes; sin embargo, los calificadores mostraron menos acuerdo en los ítems. Las correlaciones Pearson entre los calificadores fueron > 0.85. Se discuten el impacto de la fiabilidad y el nuevo método de calificación del Bender-II y las siguientes líneas de investigación con este nuevo instrumento.

Palabras clave: *Visomotor, Evaluación, Niños, Bender, Confiabilidad.*

Correspondencia:
César A. Merino Soto. Dirección Postal: Calle Filiberto Romero 430, Chorrillos – Lima 9, Perú. Email: sikayax@yahoo.com.ar
Nota del autor: la presente investigación ha sido posible por el apoyo del Instituto de Investigación de Psicología de la Universidad de San Martín de Porres. Se agradece la participación de los niños y personal educativo que aceptaron participar en el estudio.

INTER-SCORER REALIBILITY OF THE BENDER GESTALT TEST II AMONG AN INDEPENDENT SAMPLE

ABSTRACT

This paper examines inter-scorer agreement with regard to the new version of the Bender Visual-Motor Gestalt Test 2nd Edition (Bender-II). This new version includes significant changes in scoring and administration processes, and psychometric studies in Spanish language have not been published yet. Ours is among the first studies to analyze measurement error of Bender-II scores among an independent sample. The sample was made up of first and second grade primary school children from a private school, and 4 scorer students from a private university. We analyzed the level of inter-scorer agreement (intraclass correlation and Pearson) and internal consistency (Cronbach alpha). High coefficients were obtained for both the internal consistency (Cronbach alpha> 0.80) and for inter-scorer agreement in the total sample (ICC> 0.85) and in paired comparisons (ICC> 0.72) of the scores. Nevertheless, scorers showed less agreement on the items. Pearson correlations between scorers were > 0.85. We discuss the impact of reliability and the new Bender-II scoring method, as well as subsequent lines of research with this new instrument.

Key Word: *Visual-motor; Assessment, Psychometrics, Benderreliability.*

1. INTRODUCCIÓN

Una revisión informal de los silabo de pruebas psicológicas en varias universidades hispanas podría sugerir que, con respecto a la segunda versión del Test Gestáltico Visomotor de Bender (Bender-II, Brannigan & Decker, 2003), habría una resistencia al cambio cuando aparecen materiales nuevos. El Bender-II es nueva versión del Test Gestáltico de Bender (TGB, Bender; 1938) que ha sido introducido con un cambio importante en la estructura y funcionamiento de esta tradicional prueba, y de la actualmente han aparecido pocas publicaciones independientes en revistas anglosajonas (Decker, Allen, & Choca, 2006; Decker, 2007; Volker et al., 2009). Entre las publicaciones científicas internacionales en todos los idiomas, aún no se han reportado estudios que extiendan los correlatos del Bender – II en muestras independientes. En el habla hispana, no hay estudios publicados sobre sus propiedades psicométricas, en grupos diferentes a la muestra de estandarización americana (Brannigan & Decker, 2003), excepto el proveniente de un evento académico local en Puerto Rico, en que se reportaron las primeras normas hispanas para adolescentes entre 12 y 14 años (Cruz, 2008).

Han existido versiones previas que son parte de la evolución actual del Bender-II (por ejemplo, Brannigan & Brunner, 2002), y en que se examinaron aspectos psicométricos como la varianza error, que es un aspecto relacionado con la fiabilidad. Para un instrumento como el Bender-II, que incorpora subjetividad en la obtención de los puntajes, un método apropiado para obtener evidencias de fiabilidad es el método del acuerdo intercalificadores, que evalúa la variabilidad en el uso de algún sistema de calificación (Feldt & Brennan, 1989). Conocer el nivel de acuerdo entre calificadores es importante para poner límites realistas a sus resultados, y reportarlo es una práctica recomendada

para las buenas prácticas en medición psicológica y educativa (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999), y más aún en una prueba no verbal que tendría una potencial utilidad en la evaluación multicultural (Goupaul-McNicol, S. –A., 1997).

La obtención de la fiabilidad de las mediciones tiene una gran importancia en los dominios científicos y aplicados, pues la identificación de la variabilidad de las puntuaciones debido a fuentes de error permite conocer la precisión del estatus del sujeto en el constructo medido; y en un terreno profesional, ayudaría a determinar si la variabilidad en el puntaje tener significancia clínica (Jacobson, Follette & Revenstorf, 1984).

Considerando este aspecto psicométrico, casi toda la información sobre el acuerdo intercalificadores para el TGB proviene de investigaciones no hispanas. Cuando son comparados con algunas normas cualitativas para establecer el nivel de fiabilidad (por ejemplo, Cicchetti, 1994; Nunnally & Bernstein, 1995), se han hallado niveles satisfactorios de acuerdo inter-calificadores en varios de los sistemas de calificación para el TGB (Andert, 1976; Aylward & Smidth, 1986; Brannigan y Brunner, 2002; Hustak, Dinning & Watkins, 1980; Köppitz, 1975; Parsons & Weinberg, 1993; Sisto, Noronha & Santos, 2005; Swenson & Hill, 1990; Rae & Hyland, 2001). Con el nuevo Bender-II, se hallaron coeficientes desde 0.80 hasta 0.96 (Brannigan & Decker, 2006), que sugiere que la varianza de error proveniente de los calificadores es baja. Aunque los autores del Bender-II no declaran específicamente qué coeficientes usaron, se podría asumir que usaron correlaciones Pearson. Este coeficiente, sin embargo, únicamente cuantifica la monoticidad de las relaciones lineales (Cone, 1999; Esquivel et al., 2006), y por lo tanto, evalúa la consistencia de los calificadores pero no específicamente el acuerdo. Un método apropiado es usar estimaciones basadas en el análisis de varianza, específicamente correlaciones intraclase (Shrout & Fleiss, 1979; Nunnally & Bernstein, 1995; Cone, 1999) o mediante métodos que revelen la magnitud de las diferencias. Con el Bender – II, en una parte de la muestra de estandarización, los autores reportaron estimaciones de consistencia intercalificadores en dos estudios, que resultaron en coeficientes mayores a 0.80, indicando un excelente consenso en el uso del Sistema de Calificación Global; sin embargo, estas estimaciones podrían no ser apropiadas debido que el método estadístico usado.

Considerando lo anterior, el objetivo del presente artículo es presentar los resultados preliminares en la investigación de las propiedades psicométricas del Bender-II, específicamente de dos aspectos de la fiabilidad: el acuerdo inter-calificadores y la consistencia interna de sus puntajes de Copia y Recuerdo. Este aspecto investigado se refiere específicamente al monto de error debido la variabilidad de las respuestas y de los calificadores, ésta última más relevante en el Bender-II, ya que requiere un grado de juicio y discernimiento en la asignación de los puntajes.

2. MÉTODO

2.1 Participantes

La muestra de niños provino de un colegio privado de educación regular ubicado en Lima Metropolitana. Fueron 36 niños (55% varones), entre primero y segundo grado

de primaria, y dentro de los 6 años y 7 años de edad. El colegio es pequeño (15 niños o menos por aula), no es selectivo respecto a la matrícula y funciona de acuerdo a las normas del contenido curricular del gobierno peruano. El director y los padres de familia de los niños autorizaron la aplicación de la prueba para fines de investigación.

Por otro lado, los calificadores (identificadas como A, B, C y D) fueron tres estudiantes de psicología, de una universidad privada de Lima Metropolitana; tenían previo conocimiento en cursos de medición psicológica y en la aplicación y calificación de pruebas de desarrollo infantil, pero no de pruebas de habilidad visomotora.

2.2. Instrumento

Test Gestáltico Visomotor de Bender, 2da. Versión (Brannigan & Decker, 2003). Esta es la nueva versión que ha sido diseñado para sujetos desde 4 a 85 años de edad, que evalúa el funcionamiento visomotor y otros aspectos del funcionamiento cognitivo asociado a la visomotricidad. Contiene 16 láminas con diseños diferentes en cada una; los nuevos diseños se crearon para amplíar el escalamiento de los puntajes. Los niños menores de 8 años resuelven los 9 diseños originales más 4 nuevo diseños, mientras que los de 8 años a más, 3 diseños adicionales a los ya existentes. El nuevo Bender-II tiene dos fases: Copia y Recuerdo, más dos pruebas suplementarias que evalúan la motricidad fina y la percepción visual. En la fase de Copia se le pide que reproduzca todos los diseños presentados una por una; en la fase Recuerdo, se requiere que al evaluado dibuje todos los diseños que pueda recordar. Complementariamente, se le solicita resolver ítems de coordinación motora y discriminación visual. Para calificar los diseños en la fase Copia y Recuerdo, se usa el *Sistema de Calificación Global* (SCG), un método un método intuitivo y continuo para medir el cambio de la calidad de los diseños reproducidos; cada diseño se califica desde 0 (ausencia de forma en el dibujo) y 4 (dibujo reproducido casi perfecto). El manual presenta ejemplos de cada categoría de puntuación. El Bender-II desarrolló sus aspectos normativos y psicométricos desde una muestra de más de 4000 personas estratificadas por etnia, educación y estatus socioeconómico, seleccionados de acuerdo con el censo de la población americana del año 2000 (Brannigan & Decker, 2003). La información de todos los aspectos de validez explorados con medidas de inteligencia, visomotricidad, rendimiento escolar, y comparaciones entre grupos de diferencias condiciones, muestran una buena capacidad discriminativa y correlaciones teóricamente respaldadas.

2.3. Procedimiento

La administración del Bender-II se hizo manteniendo las instrucciones del manual respecto al orden de las fases de evaluación, de tal manera que se procuró satisfacer las condiciones mínimas de administración estandarizada en la interacción individualizada (Lee, Reynolds y Willson, 2003; Bracken, 2007); esto significa, la creación del clima de confianza, lo apropiado del ambiente, la mantención de la motivación, la introducción de pausas necesarias en respuesta al posible cansancio del niño, y la minimización de la ocurrencia y el impacto de eventos periféricos durante la evaluación.. Las evaluaciones se hicieron en una sesión individual por cada niño y en un aula, controlándo

con las mínimas distracciones posibles. Los examinadores fueron los mismos que calificarían posteriormente los diseños reproducidos. Por otro lado, la calificación se hizo después de un entrenamiento de dos sesiones, en que se enfatizó la evaluación global del diseño y cómo resolver las dudas en la elección de dos puntuaciones consecutivos (por ejemplo entre el puntaje 2 y 3).

Los análisis estadísticos consistirán en la evaluación de dos fuentes de error relacionados con: la consistencia interna y el acuerdo entre calificadores. Se aplicará el coeficiente α (Cronbach, 1951) asumiendo un modelo equivalente tau entre los ítems; también se reportará la correlación inter-ítem promedio y el resumen de las correlaciones ítem-test. Para justificar el uso del coeficiente α, se verificó la unidimensionalidad mediante la interpretación del primer autovalor extraído y la tasa entre esta y el segundo autovalor, que debería ser tres o cuatros veces mayor (Hattie, 1985). Los autovalores y la varianza retenida por el primer componente (entre parénetesis) para las cuatro calificadoras A, B, C y D, fueron 6.05 (46%), 4.52 (34%), 5.15 (39%) y 6.19 (47%), respectivamente; cuando estos valores fueron comparados con los 2dos autovalores, las tasas fueron mayores a 3.5. Estos resultados sugieren suficiente unidimensionalidad de los puntajes del Bender-II en esta muestra de participantes.

Para estimar el acuerdo inter-calificadores se aplicará un modelo del análisis de varianza, con el coeficiente de correlación intraclase, *ICC* (Shrout & Fleiss, 1979), modelo 2; este modelo de ICC asume que los calificadores son seleccionados aleatoriamente de alguna población de calificadores potenciales y cada calificador evalúa a cada examinado; este es el modelo de efectos aleatorios de dos vías y cubre mayormente

TABLA 1
ESTADÍSTICOS BÁSICOS Y CONSISTENCIA INTERNA DE LOS PUNTAJES COPIA Y RECUERDO DEL BENDER-II

Puntaje	M	DE	Min	Max	Sim.[a]	Curt.[b]	a		R_{ii}	R_{itc}		
							a_{se}[c]	a_{in}		Pro	Min	Max
Copia												
J	31.78	7.650	14	45	-0.69	0.53	0.89	0.90	0.40	0.60	0.24	0.76
P	29.92	6.991	14	44	-0.38	0.53	0.83	0.84	0.27	0.48	0.22	0.64
E	31.80	7.70	16	46	-0.45	-0.19	0.86	0.87	0.32	0.53	0.21	0.76
D	34.44	7.129	17	46	-1.45	0.77	0.90	0.91	0.41	0.61	0.25	0.79
Recuerdo												
J	9.19	5.544	0	24	1.18	-0.61	--		--	--	--	--
P	8.81	6.140	0	26	1.69	-0.44	--		--	--	--	--
E	10.22	5.718	0	23	0.67	-1.02	--		--	--	--	--
D	11.61	6.796	0	30	1.52	-0.08	--		--	--	--	--

[a]: error estándar = 0.393; [b]: error estándar = 0.768; [c]: α de Cronbach sesgado; [d]: α de Cronbach insesgado (Feldt, Woodruft & Salih (1987).

las situaciones de acuerdo inter-calificadores (McGraw & Wong, 1996). Finalmente, Ciccheti (1994) declara cuatro niveles de evaluación cualitativa aplicable al acuerdo inter-calificadores: < 0.40 = pobre, 0.40 – 0.59 = aceptable, 0.60 – 0.74 = bueno, > 0.74 = excelente. Para la consistencia interna, Ciccheti considera que <0.70 es inaceptable, 0.71 a 0.79 es aceptable, 0.80 a 0.89 es bueno, y >= 0.90 es excelente.

3. RESULTADOS

Los resultados descriptivos y de la consistencia interna se muestran en la Tabla 1, y los resultados del acuerdo se presentan en la Tabla 2.

Consistencia interna. La consistencia interna de los puntajes de las calificadoras fue más de 0.80 (Tabla 1), y pueden definirse como buenos y excelentes niveles consistencia interna. Debido al error estándar afectado por el tamaño de la muestra, se aplicaron intervalos de confianza (I.C.) y se corrigió el sesgo de estimación por muestras pequeñas al coeficiente **a** (**a**$_{in}$). Ambos procedimiento vienen de Feldt, Woodruft & Salih (1987). Para completar la información, también se presenta el coeficiente **a** sin corregir o sesgado (**a**$_{se}$) Para las calificadoras A, B, C y D, los I.C. al 95% fueron respectivamente:

TABLA 2
CORRELACIONES PEARSON E ÍNDICES DE ACUERDO (ICC CON INTERVALOS DE CONFIANZA 95%) ENTRE LOS CALIFICADORES EN EL PUNTAJE COPIA Y RECUERDO

Copia				
Total (ICC)	0.85 [0.70, 0.93]			
Pareados				
	J[b] .	P .	E .	D .
J[a]	1	0.90	0.92	0.90
P	0.87 [0.70, 0.94]	1	0.94	0.87
E	0.93 [0.86, 0.96]	0.91 [0.69, 0.96]	1	0.90
D	0.85 [0.47, 0.94]	0.73 [-0.0, 0.91]	0.84 [0.49, 0.93]	1
Recuerdo				
Total (ICC)	0.85 [0.74, 0.92]			
Pareados				
	J[b] .	P .	E .	D .
J[a, c]	1	0.88	0.86	0.86
P	0.88 [0.78, 0.93]	1	0.91	0.96
E	0.85 [0.72, 0.92]	0.91 [0.69, 0.96]	1	0.94
D	0.79 [0.45, 0.91]	0.73 [-0.0, 0.91]	0.84 [0.49, 0.93]	1

[a]: En el triángulo inferior, se informa de las correlaciones ICC; [b]: en el triángulo superior, se informa de la correlación de Pearson; [c]: las leras J, P, E y D identifican a las calificadoras participantes del estudio.

[0.84, 0.94], [0.75, 0.91], [0.80, 0.92] y [0.86, 0.95]. Aplicando los criterios de cualificación (Ciccheti, 1994), los coeficientes obtenidos pueden ser considerados entre los niveles de "aceptable" y "bueno" para la presente muestra. Esta variación depende del tamaño de los intervalos de confianza y del tamaño muestral del estudio. Por otro lado, se puede observar que los otros índices de consistencia interna (Tabla 1), la correlación inter-ítem promedio (R_{ii}) y la discriminación de los ítems (correlación ítem-test corregida, R_{itc}) también muestran magnitudes apropiadas para considerar al Bender-II como una medida suficientemente homogénea, aunque no tan altos como para indicar que contiene ítems redundantes (Boyle, 1991).

Acuerdo intercalificadores. Para los puntajes de Copia y Recuerdo, la magnitud de los índices de acuerdo (ICC) prácticamente fue similar entre las calificadoras (Tabla 2). Excepto una de calificadoras que mostró más variabilidad respecto a las otras (ICC debajo de 0.70), los niveles de acuerdo entre las calificadoras estuvo alrededor de 0.86, y los intervalos de confianza indican que el acuerdo puede tener un rango de bueno hasta excelente concordancia (Ciccheti, 1994). La consistencia de las calificaciones, estimadas mediante las correlaciones Pearson (situados en la zona superior derecha de la Tabla 2), también fue elevada, y muy similares a los coeficientes ICC. Considerando el acuerdo en los ítems (Tabla 3), se obtuvieron coeficientes bajos relativos a las estimaciones ICC en el puntaje total. No parece observarse un claro patrón de discrepancias entre las calificadoras en los ítems. Excepto el diseño 2, el resto de los tres nuevos diseños muestran magnitudes aceptables o buenos niveles de acuerdo. El diseño 11, aunque no es nuevo, ha sido sensible mayores discrepancias en las calificadoras.

TABLA 3
ACUERDO ICC PARA LOS ÍTEMS

Diseño	ICC (2, 4)	I.C. 95%
1	0.63	[0.48, 0.76]
2	0.47	[0.30, 0.64]
3	0.52	[0.35, 0.68]
4	0.56	[0.40, 0.71]
5	0.73	[0.55, 0.85]
6	0.74	[0.62, 0.84]
7	0.55	[0.38, 0.70]
8	0.68	[0.53, 0.80]
9	0.67	[0.53, 0.79]
10	0.64	[0.49, 0.77]
11	0.49	[0.32, 0.65]
12	0.70	[0.57, 0.82]
13	0.74	[0.61, 0.85]

4. DISCUSIÓN

El objetivo del estudio fue presentar resultados de la fiabilidad de los puntajes del Bender-II, en un grupo de escolares entre primer y segundo grado de primaria. Los estudios sobre las propiedades psicométricas del Bender-II aún son escasos en habla anglosajona, y menos aún en hispana, y los resultados encontrados forman parte de las primeras informaciones empíricas sobre este instrumento. El énfasis del presente análisis fue el acuerdo intercalificadores y la consistencia interna. Se hallaron excelentes niveles de fiabilidad en la consistencia interna y en el acuerdo entre calificadores; estas estimaciones se mantuvieron elevadas entre las cuatro calificadoras considerando el puntaje total, pero el nivel de acuerdo fue relativamente bajo en los ítems.

Los resultados mostraron que la consistencia interna del puntaje es elevada, y puede clasificarse como de buen o excelente nivel de fiabilidad. Relacionado con el rendimiento del niño, la elevada consistencia interna sugiere que las respuestas de los niños pueden ser parecidas en todos los ítems, y que las diferencias grandes diferencias en los puntajes en un diseño y otro son menos frecuentes. Para incluir una mejor perspectiva de los resultados, se calculó la consistencia interna de los puntajes provenientes de las cuatro calificadoras. Excepto una de ellas, la variabilidad del coeficiente α generalmente estuvo entre los niveles *excelente* y *bueno* (Cichetti, 1994), indicando un reducido impacto de la varianza de error en los puntajes del Bender-II. Debido que el tamaño de la muestra del estudio podría producir una subestimación del coeficiente α, se aplicó un ajuste para reducir el sesgo (Feldt, et al., 1987). Los coeficientes ajustados produjeron apenas un leve incremento, pero serán más precisos para informar cuantitativamente del grado de error en los puntajes del Bender-II. Los niveles reportados de fiabilidad pueden servir como referencias del grado de varianza de error que puede minimizarse durante el diseño de investigación cuantitativa (Gil, 2003).

Considerando la variabilidad entre los examinadores, ésta fue leve, y por lo tanto el monto de error es bajo. Las estimaciones de acuerdo obtenidas en el manual fueron más elevadas, pero como se afirmó antes, sus correlaciones fueron posiblemente coeficientes Pearson. Las correlaciones Pearson obtenidas en la presente muestra fueron más elevadas que las ICC, y dado que sí son comparables con los coeficientes Pearson del manual, entonces se puede concluir que la consistencia obtenida entre las cuatro calificadoras fue similar a las reportadas en el manual.

Respecto al tamaño muestral, los intervalos de confianza generados proporcionaron la información necesaria para evaluar el rango de magnitudes en que el acuerdo puede ocurrir en la población. Excepto el acuerdo entre un par de calificadoras, los mínimos niveles posibles generalmente estuvieron en el nivel de *buen* acuerdo (Cicchetti, 1994). La magnitud de los ICC fue algo más elevada en las reproducciones de Copia, y parece que el grado de distorsión en las reproducciones de la fase Recuerdo sea de mayor variabilidad y severidad, considerando que los diseños mal recordados tendían a ser una mezcla de varias figuras. El manual no indica cómo resolver una situación así durante la calificación (mezcla de figuras), y ello traería como consecuencia que el calificador desarrolle criterios personales para evaluar la exactitud del diseño, más aún si no hay mucha experiencia en el uso de pruebas para niños, y en pruebas de visomotricidad, específicamente.

Los resultados presentados aportan en la generación de información psicométrica del Bender-II en grupos diferentes a la muestra de estandarización, indicando que se pueden obtener niveles de acuerdo intercalificadores relativamente elevados luego de un entrenamiento de por lo menos dos sesiones. El manual informa que usuarios sin entrenamiento en pruebas visomotoras y con solo una práctica autodidacta pueden obtener puntajes similares a calificadores experimentados (Brannigan y Decker, 2003); esto puede señalar la relativa facilidad para aplicar el SCG. En la presente experiencia de entrenamiento y monitoreo de las calificadoras, se enfatizó la comprensión de una de las reglas básicas para aplicar correctamente el SCG: la asignación del puntaje adyacente inferior cuando hay dudas en la calificación. Este aspecto del SCG presentó algunas complicaciones en el aprendizaje inicial, pues fue aprendido como una heurística aplicable sin un juicioso examen del diseño y como consecuencia, produjo puntajes más conservadores. Esto puede significar que disminuyó la orientación a percibir la gestalt desde un criterio evolutivo, y más bien las calificadoras pudieron centrarse en el acabado artístico o en los tipos de desvíos específicos con respecto al diseño original.

Una de las limitaciones del presente estudio es el tamaño de la muestra, pues es una situación que produce un elevada inestabilidad de las estimaciones estadísticas (Cohen, 2007). Más aún, cuando se ha recomendado que el tamaño muestra mínimo recomendado para una más precisa estimación de la fiabilidad es 300 o más (Charter, 1999). Sin embargo, recientes artículos metodológicos indican que hay condiciones que podrían dar un suficiente respaldo para obtener un coeficiente α aproximadamente robusto en un tamaño muestral tan pequeño como, por ejemplo, 30 sujetos (Charter, 2008; Yurgudúl, 2008). Una de estas condiciones es el tamaño de primer autovalor extraido (Yurgudúl, 2008), y tal magnitud se halló en cada una de los autovalores proveniente de los ítems calificados por las calificadoras. Por lo tanto, se puede considerar que las estimaciones de fiabilidad obtenidas son aceptablemente representativas. También se debe considerar que la situación de realizar un estudio en una muestra pequeña no se diferencia de la cantidad de sujetos que el profesional afronta en la práctica de consultorio, y en que se deben tomar decisiones individuales o de pequeños grupos con la información de fiabilidad obtenida. En este punto, el Bender-II puede alcanzar buenas características psicométricas cuando se mantiene un cuidadoso estudio del manual y aplicación correcta.

Por otro lado, la muestra de calificadores fueron estudiantes, y esta situación debería prevenir al lector sobre generalizaciones de los resultados hacia otros grupos de usuarios; por ejemplo, psicólogos graduados o en práctica profesional. Es recomendable extender una investigación sobre el acuerdo entre calificadores con experiencia profesional, en condiciones de aprendizaje autodidacta versus aprendizaje monitoreado, y evaluar el grado de acuerdo en varios niveles de puntuación. Los resultados obtenidos con este diseño podrían dar diferentes estimaciones del acuerdo, y mejorar la comprensión de las características de la variabilidad de las calificaciones en el Bender-II. Comprobaciones como estas serían fundamentales para un aprendizaje y uso autónomo por parte del usuario, y para disminuir el riesgo de falsos positivos o falsos negativos en el diagnóstico de las habilidades visomotororas.

Dos aspectos no resueltos surgen del presente estudio: en primer lugar, la estimación de la consistencia interna de los puntajes puede ser el límite inferior de la verdadera

fiabilidad si el modelo de medición de los diseños no se ajusta al modelo equivalente tau (Feldt & Brennan, 1989; Feldt & Charter, 2003). Una comprobación del modelo de medición en una muestra más grande podrá verificar si lo obtenido aquí es replicable, y si la violación de este supuesto estadístico. En segundo y último lugar, se ha evaluado la variabilidad entre calificadores, pero no la variabilidad intra-calificador, o también denominado *deterioro de la fiabilidad* (Torres & Perera, 2009). Este aspecto debería evaluarse en un diseño de medidas repetidas para completar la evaluación de la fiabilidad del nuevo Bender-II. Por el momento, los resultados presentados parecen un buen incentivo para considerar el uso del Bender-II en la práctica profesional y la investigación en el mundo hispano.

5. REFERENCIAS BIBLIOGRÁFICAS

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing*. Washington DC: American Psychological Association.

Aylward, E. H., & Schmidt, S. (1986). An examination of three test of visual-motor integration. *Journal of Learning Disabilities, 19*(6), 328-330.

Bender, L. (1938). A visual-motor gestalt test and its clinical use. *American Orthopsychiatric Association Research Monographs*, No. 3.

Boyle, G. J. (1991). Does item homogeneity indicate internal consistency or item redundancy in psychometric scales? *Personality and Individual Differences*, *12*(3), 291-294.

Bracken, B. (2007). Creating the optimal preschool testing situation. In B. Bracken & R. J. Nagle (Eds.) *Psychoeducational assessment of preschool children* (pp. 137-153). Mahwah, N.J.: Lawrence Erlbaum Associates.

Brannigan, G. G., & Brunner, N. A. (2002). *Guide to the qualitative scoring system for the Modified Version of the Bender-Gestalt Test.* Springfield, IL: Thomas.

Charter, R. A. (1999). Simple size requeriments for precise estimates of reliability, generalizability, and validity coefficients. *Journal of Clinical and Experimental Neuropsychology*, *21*, 559-566.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*, 284-290.

Cohen, B. H. (2007). *Explaining Psychological Statistics (3rd ed.)*. New York: John Wiley & Sons.

Cone, J. D. (1999). Observational assessment: Measure development and research issues. In P. C. Kendall, J. N. Burcher, & G. N. Holmbeck, *Handbook of Research Methods in Clinical Psychology (2nd ed.),* (pp. 183 - 223). NY: John Wiley & Sons.

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

Cruz, D. (Mayo, 2008). Desarrollo de Normas para la Prueba de Desarrollo Viso- Motor Bender II en Estudiantes Puertorriqueños de 12, 13 y 14 años. Ponencia presentada en el Congreso de Medición: Innovación, tecnología y nuevas prácticas en la psicometría. Asociación de Psicología de Puerto Rico, Universidad Central de Bayamón.

Cummings, J. A., Hoida, J. A., Machek G. R., & Nelson, J. M. (2003). Visual-motor assessment of children. In C. R. Reynolds & R. W. Kamphaus (Eds.) *Handbook of psychological and educational assessment of children: intelligence, aptitude, and achievement (2nd. Ed.,* pp. 498-518). New York: Guilford Press.

Decker, S. L. (2007). Measuring growth and decline in visual-motor processes using the Bender-Gestalt II. *Psychoeducational Assessment, 26*(1), 3-15.

Decker, S. L., Allen, R., & Choca, J. P. (2006). Construct validity of the Bender-Gestalt II: Comparison with Weschler Inteligence Scale for Children-III. *Perceptual and Motor Skills, 102*, 133-141.

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.

Fedlt, L. S., Woodruft, D. J., & Salih, F. A. (1987). Statistical inference for coefficient alpha. *Applied Psychological Measurement*, *11*(1), 93-103.

Feldt, L. S. & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105-146) Washington, DC: American Counsil on Education.

Feldt, L. S., & Charter, R. A. (2003). Estimating the reliability of a test split into two parts of equal or unequal length. *Psychological Methods*, *8*, 102-109.

Gil, J. (2003). La estadística en la investigación educativa. *Revista de Investigación Educativa*, *21*(1), 231-248.

Goupaul-McNicol, S. –A. (1997). *A multiciultural/multimodal/multisystems approach to working with culturally differents families*. Wesport, CT: Praeger.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9,* 139-164.

Hustak, T. L, Dinning, W. D., & Andert, J. N. (1976). Reliability of the Koppitz scoring system for the Bender Gestalt Test. *Journal of Clinical Psychology, 32*(2):468-9.

Jacobson, N. S., Follette, W. C., & Revenstorf, D. (1984). Psychotherapy outcome research: Methods for reporting variability and evaluating clinical significance. *Behavior Therapy, 15,* 336–352.

Koppitz. E. M. (1975). *The Bender-Gestalt Test for young children: II Research and application, 1963-1973*. New York: Grune & Stratton.

Lee, D., Reynolds, C. R., & Willson, V. L. (2003). Standardized test administration: Why bother? *Journal of Forensic Neuropsychology*, *3*, 55-81.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*(1), 30-46.

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

Parsons, L., & Weinberg, S. L. (1993).The Sugar Scoring System for the Bender-Gestalt. *Perceptual and Motor Skills, 77*, 883-893.

Shrout, P. E. & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*,420-428.

Sisto, F. F., Noronha, A. P. P., & Santos, A. A. A. (2005). *Bender - Sistema de Pontuação Gradual B-SPG*. Vetor Editora: São Paulo.

Svensson, P.W., & Hill, M.A. (1990). Interrater reliability of the Koppitz developmental scoring method in the clinical evaluation of the single case. *Perceptual and Motor Skills,* 70, 615-623.

Torres, J., & Perera, V. (2009). Cálculo de la fiabilidad y concordancia entre calificadores de un sistema de categorías para el estudio del foro online en e-learning. *Revista de Investigación Educativa*, *27*(1), 89-103.

Volker, M. A., Lopata, C., Vujnovic, R. K., Smerbeck, A. M., Toomery, J. A., Rodgers, J. D., Schiavo, A., & Thomeer, M. L. (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.

Watkins, E. (1980) *Sistema de Puntuación de Watkins para el Test Gestáltico Visomotor*. Buenos Aires: Panamericana.

Yurgudúl, H. (2008). Minimum sample size for Cronbach's coefficient alpha: A Monte Carlo study. *Hacettepe Üniversitesi Journal of Education*, *35*, 397-405.

Fecha de recepción: 20 de febrero de 2011.
Fecha de revisión: 20 de febrero de 2011.
Fecha de aceptación: 28 de agosto de 2011.