Acuerdo intercalificadores intragrupo en el Test Gestáltico Visomotor de Bender, segunda versión (Bender - II)*

Within-group inter-rater agreement for the Bender Gestalt Visualmomor Test, second version (Bender - II)

Recibido: 30 de marzo de 2012| Aceptado: 29 de febrero de 2016

César Merino Soto**
Instituto de Investigación de Psicología
de la Universidad de San Martín de Porres, Perú
Ryan Allen***
John Carroll University, Estados Unidos
Scott Decker****
University of South Carolina

Resumen
La presente investigación evalúa el acuerdo intercalificadores del Test Gestáltico Visomotor de Bender, segunda versión (Bender - II). Esta nueva versión ha mostrado propiedades psicométricas satisfactorias en la muestra de estandarización y en posteriores estudios anglosajones, pero no hay referencias sobre sus propiedades en población hispanos. Los participantes de este estudio fueron 173 niños preescolares entre cuatro y cinco años de edad; los calificadores fueron estudiantes de psicología. Para evaluar la consistencia y el acuerdo entre los calificadores, se aplicaron coeficientes de correlación Pearson, así como correlaciones intraclase (modelo 2). Se calcularon estos coeficientes para la muestra total y para el grupo superior e inferior de rendimiento visomotor, aplicando un ajuste por restricción del rango. Los resultados indican elevadas confiabilidades y acuerdo entre calificadores en la muestra total; pero en los grupos inferior y superior de rendimiento, estos fueron variables y algunos bajos, interaccionando con la edad. Estos resultados ocurrieron para los puntajes de "copia" y de "recuerdo". Se discute el efecto de la variabilidad entre calificadores, especialmente su impacto en las decisiones clínicas sobre desempeños bajos y altos.
Palabras clave
evaluación; confiabilidad; psicometría; niños; visomotor

Abstract
The present research evaluates the agreement inter-scorers for the Bender Gestalt Test Bender, 2nd version (Bender - II). This new version has shown satisfactory psychometric properties in the sample for standardization, and in further studies English samples, but no references in Hispanic population. The participants in this study were 173 preschool children between 4 and 5 years of age; the scorers were students of psychology. To evaluate the internal consistency and agreement inter-raters, we applied Pearson correlation coefficients and intraclass correlations (model 2). These coefficients were calculated for the total sample and for the upper and lower group sample in visual-motor performance, but with adjustment by the restriction of range. The results for the Copy and Remember scores indicate high internal consistency and agreement between raters in the total sample, but low and variable in the lower and upper groups, interacting with age. We discuss

doi: 10.11144/Javeriana.upsy15-2.aiit

Para citar este artículo: Merino Soto, C., Allen, R., & Decker, S. (2016). Acuerdo intercalificadores intragrupo en el Test Gestáltico Visomotor de Bender, segunda versión (Bender - II). *Universitas Psychologica, 15*(2), 163-172. http://dx.doi.org/10.11144/Javeriana.upsy15-2.aiit

* Artículo de investigación.

** Correo electrónico: sikayax@yahoo.com.ar

*** Department of Education & Allied Studies. Correo electrónico: rallen@jcu.edu

**** Department of Psychology, Barnwell College. Correo electrónico: sdecker@gsu.edu

the effect of inter-rater variability, especially its impact on clinical decisions about low and high performance.
Keywords
assessment; reliability; psychometry; children; visualmotor

Desde que Lauretta Bender creó el Test Gestáltico Visomotor de Bender -TGB- (Bender, 1938) y publicara un manual años después (en 1946), éste continuó como una de las medidas más populares y utilizadas por profesionales de las salud mental (Archer & Newsom, 2000; Piotrowski, 1995; Sullivan & Bowden, 1997). El TGB ha sido descrito como una medida cognitiva y del desarrollo de habilidades visomotoras, visuales-constructivas y perceptomotoras (Lacks, 1999). Algunos clínicos lo han usado para evaluar atributos de personalidad, aunque no es un uso ampliamente reconocido (Hutt, 1969, 1985; Perticone, 1998; Raphael & Reichenberg, 1992). La generación de diferentes sistemas de calificación ha llevado a que muchos investigadores realizaran estudios sobre su validez y confiabilidad, incrementándose esto por los años 90 (Cummings, Hoida, Machek, & Nelson, 2003). En Estados Unidos, Watkins, Campbell, Nieberding, & Hallmark (1995) reportaron que el TGB es una de diez medidas preferidas entre los psicólogos y su comunalidad con otras pruebas para el mismo fin se origina en el tipo de tarea solicitado al examinado, es decir, en el copiado de figuras geométricas (Cummings et al., 2003).

Aun con su popularidad, muchos de los sistemas de calificación para el TGB tuvieron problemas de validez y confiabilidad. En el 2003, Brannigan & Decker revisaron el TGB original, adicionaron siete láminas y abordaron sus debilidades psicométricas (es decir, limitada muestra normativa y procedimientos de administración estandarizados). Esta nueva versión del TGB (la segunda versión: Bender – II) se califica usando una versión adaptada del Sistema de calificación cualitativa (SCC) para el Test Gestáltico Visomotor de Bender – Modificado (Brannigan & Bruner, 2002) el cual, se diseñó para evaluar la calidad global de las reproducciones en una versión modificada del TGB (contiene seis láminas). Este sistema ha demostrado satisfactorias cualidades psicométricas en niños hispanos (Merino & Benites, 2011; Merino, 2009, 2010, 2011) En el Bender – II, el nuevo método de calificación se denomina Sistema de calificación global; este sistema de calificación se centra en que la calidad del diseño reproducido por la persona evaluada sea igual o mejor que el ejemplo dado en el manual, otorgándose una calificación ordinal a cada ítem (lámina).

El Bender – II se puede aplicar a personas desde los cuatro a los 80 años y actualmente es la siguiente generación del tradicional TGB publicado hace más de 50 años por L. Bender (1938, 1946), cuya popularidad entre los psicólogos creció exponencialmente con el advenimiento de los trabajos de Koppitz en niños (1963, 1975) y que aún parece ser constante hasta la década pasada (Watkins et al, 1995). Además de que el Bender-II tenga la ventaja de ser un instrumento único para un amplio rango de edad, hay aspectos estructurales y funcionales importantes que lo convierte en un instrumento originalmente nuevo. Por ejemplo, actualmente contiene cuatro fases de evaluación: el primero de ellos es el copiado de los diseños; luego, el recuerdo de los diseños copiados; continuando con tareas de control motor, y finalmente con una prueba de discriminación perceptual visual. Estos últimos componentes estructurales amplían los usos del TGB dentro de una evaluación integrada para la detección de problemas de memoria, visomotricidad y el despistaje de problemas en la motricidad y percepción visual. Las normas de interpretación usan diferentes puntajes estandarizados. El TGB se generó desde una muestra grande y representativa de sujetos entre 4 y 80 años; sin embargo, estas normas se desarrollaron únicamente en Estados Unidos, así como sus propiedades psicométricas. En los recientes estudios desde los que se derivan sus propiedades de validez de constructo y confiabilidad (Decker, 2007; Volker et al., 2010; Merino, 2012), se puede resaltar su robustez psicométrica como una medida de las habilidades visomotoras en niños, adultos y en poblaciones especiales. Estas características han hecho que el Bender – II sea una de las pocas medidas de habilidad visomotora designada como una herramienta bien establecida

para la práctica evaluativa pediátrica, de acuerdo a la revisión crítica efectuada por la *Evidence-Based Assessment Task Force Workgroup for the Society of Pediatric Psychology, Division 54* (Campbell, Brown, Cavanagh, Vess, & Segall, 2008). Entre las publicaciones científicas internacionales en idioma no anglosajón, aún no se han reportado estudios que extiendan los correlatos del Bender – II en muestras independientes.

En el habla hispana hay escasos estudios publicados sobre las propiedades psicométricas en muestras diferentes a la muestra de estandarización americana; estos fueron reportados para una muestra pequeña en Perú (Merino, 2012) en un evento local en Puerto Rico, donde se reportaron las primeras normas hispanas para adolescentes entre 12 y 14 años (Cruz, 2008). Sus hallazgos más significativos fueron que no hubo diferencias sustanciales entre las edades muestreadas y que las puntuaciones promedio en el copiado fueron superiores a las normas americanas. En habla hispana tampoco hay estudios publicados que repliquen el monto de varianza de error para el Bender-II, específicamente en que la fuente de error provenga de la calificación de los examinadores. El método más empleado para comprobar la confiabilidad en pruebas que exigen juicio y subjetividad es el análisis de la variabilidad de los calificadores, debido a que esta fuente de error de medición es particularmente sensible a las variaciones en el uso de algún sistema de calificación por parte de los calificadores (Feldt & Brennan, 1989). Pero en general, con las versiones anteriores, los estudios de la varianza de error en la calificación han hallado satisfactorios niveles de acuerdo entre los calificadores; estos niveles se han encontrado tanto para los métodos basados en calificaciones objetivas en las reproducciones de niños (Hustak, Dinning, & Andert, 1976; Svensson & Hill, 1990; Parsons & Weinberg, 1993; Watkins, 1980;) y adultos (Lacks, 1999),como en los sistemas de calificación global o cualitativa (Brannigan & Brunner, 2002; Pauker, 1976).

Con la versión original, Aylward & Schmidth (1986) hallaron que aunque se pueden hallar aceptables niveles de acuerdo intercalificador, la confiabilidad puede variar según el nivel de desempeño del examinado y, por lo tanto, esto nos lleva a pensar que no se podría asumir la consistencia en la clasificación de niños en el rango total del puntaje. Este aspecto parece no haber sido replicado a menudo en las investigaciones psicométricas con el TGB, sino más bien los efectos de la experiencia; por ejemplo, hay evidencias de esta experiencia entre novicios y expertos en pruebas visomotoras, estos pueden converger en elevados niveles de acuerdo (Brannigan & Decker, 2003; Lacks & Newport, 1980; Reynolds, 2007).

Para el Bender-II, el manual reporta confiabilidades interjueces desde 0.83 hasta 0.94 (Brannigan & Decker, 2003), proveniente de dos estudios. En el primero, cinco calificadores experimentados calificaron 30 protocolos y la confiabilidad promedio de las correlaciones pareadas entre las diez combinaciones posibles de calificadores fue 0.96. En el segundo estudio, dos calificadores con y sin experiencia en evaluación psicológica y en el Bender – II calificaron 60 protocolos y llegaron a un acuerdo mayor a 0.80 para las fases de "copia" y "recuerdo". Recientemente, un estudio obtuvo confiabilidades intercalificadores entre 0.91 y 0.99, en la evaluación de las reproducciones de niños autistas y normales (Volker et al., 2010). En Latinoamérica, en una muestra de escolares de primaria, Merino (2012) reportó acuerdo interexaminadores excelentes, mientras que a nivel de los ítems el acuerdo fue menor. Estos reportes sobre el Bender-II informan de un excelente consenso en el uso del Sistema de calificación global, pero estas estimaciones del impacto del error de medición no podrían ser completamente apropiadas. Aunque los autores no declaran específicamente qué coeficientes usaron, se puede asumir correctamente que usaron correlaciones Pearson. Pero estos coeficientes cuantifican la monoticidad lineal de las relaciones (Cone, 1999) y es inadecuado para identificar precisamente el grado de acuerdo (Bland & Altman, 1986). En cambio, es apropiado usar estimaciones basadas en el análisis de varianza, específicamente correlaciones intraclase (Shrout & Fleiss, 1979; Nunnally & Bernstein, 1995; Cone, 1999).

Aunque los hallazgos sobre la varianza de error reportados en el manual del Bender-II y en

la muestra peruana (Merino, 2012) han revelado que se puede lograr consistencia en diferentes grupos de calificadores y poblaciones, estos se calcularon en la muestra total de participantes. Esto no permite reconocer si el mismo nivel de acuerdo intercalificadores estimado en el rango completo de puntajes es similar en un particular nivel de desempeño, por ejemplo en los puntajes de bajo o alto rendimiento. La definición de un grupo de alto o bajo rendimiento se hizo eligiendo dos puntos de corte que separan a los grupos bajo, medio y alto (Anastasi & Urbina, 1997). En esta situación, se podría examinar si el acuerdo intercalificadoreses es igualmente aceptable en estos rangos de puntajes en que se toman decisiones importantes para los examinados (Anastasi & Urbina, 1997). La estimación de la varianza de error en los puntajes ha sido enfatizada en los reportes que incluyan mediciones conductuales, particularmente en los niveles de puntuación que influencian las decisiones sobre los evaluados (AERA, APA, & NCME, 1999).

La presente investigación tiene dos objetivos: primero se propone obtener evidencias de confiabilidad en la calificación del Bender-II en una muestra de calificadores y de participantes de diferente condición a los reportados en investigaciones previas (Brannigan & Decker, 2003; Merino, 2012; Volker et al., 2010). En segundo lugar se apunta a un objetivo más específico, que es investigar el acuerdo intercalificadores en la muestra total de estudio y en los grupos extremos bajo y alto de puntuación. Al focalizar la evaluación del acuerdo en las calificaciones del Bender-II en estos grupos, se puede verificar directamente si puede generalizarse los niveles de confiabilidad usualmente hallados en el rango completo de puntuación. Desde mucho antes se ha reconocido que el error de medición no es constante en el rango completo de puntación de una medida (Feldt & Brennan, 1989) y explorar el grado de variabilidad entre calificadores en dos regiones críticas de puntuación tiene suficiente justificación tomando en cuenta los actuales requerimientos técnicos que deben ser explorados en los instrumentos de medición (AERA et al., 1999). Los objetivos del presente trabajo también se enmarcan en la revisión hecha por Salvia, Ysseldyke, & Bolt (2009), quienes mostraron que las calificaciones en el puntaje de "copia" son menos consistentes entre los calificadores y algo más elevados para el puntaje "recuerdo". Aunque esta afirmación no fue validada con los posteriores resultados de Volker et al. (2010), no hay estudios hispanos al respecto que puedan dar un soporte adicional al nuevo Bender-II.

Método

Participantes

Fueron 173 niños de ambos sexos (83 niñas, 48%) distribuidos similarmente en las edades de cuatro (78, 45.1%) y cinco años de edad, todos procedentes de centro educativo privado de educación regular, ubicado en la zona urbana de Lima Metropolitana. Todos los niños de nuestro estudio asisten regularmente a sus clases diarias. Los niños provinieron de un ambiente familiar con las siguientes características: son de nivel socioeconómico medio y viven generalmente con ambos padres; estos tienen al menos estudios técnicos de tres años o más. Los padres principalmente son empleados en trabajos estables y las madres mayormente se desempeñan como amas de casa y algunos trabajos independientes. Cerca de la mitad de las familias conviven únicamente entre sí, mientras que el resto convive con otros familiares. Por otro lado, la mayoría de las familias residen cerca de dos kilómetros alrededor de colegio. De acuerdo a las indagaciones hechas con el director de la institución educativa, los niños no participaron en alguna intervención sistemáticamente, *ad hoc*, individual o grupal, orientada al mejoramiento de las habilidades visuales o motoras en casa o en el colegio.

Respecto a los calificadores, éstas fueron tres estudiantes de psicología de una universidad en Lima, todas ubicadas en el tercio superior de rendimiento académico. Las calificadoras contaban con experiencia en la aplicación y calificación de pruebas visomotoras, así como en pruebas de desarrollo psicomotor.

Instrumento

Test Gestáltico Visomotor de Bender, segunda versión (Brannigan & Decker, 2003). Esta es la nueva versión que ha sido diseñada para sujetos desde cuatro hasta 85 años de edad, con el objetivo de evaluar el funcionamiento visomotor y, complementariamente, otros aspectos del funcionamiento cognitivo asociado a la visomotricidad. Consta de 16 láminas de fondo blanco, cada una con un diseño diferente; los nuevos diseños añadidos amplían el escalamiento de los puntajes. Los niños menores de ocho años, resuelven los ítems originales más cuatro diseños, mientras que los de ocho años o más resuelven tres ítems 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" al sujeto se le pide que reproduzca todos los diseños presentados uno por uno; luego se le pide al evaluado que recuerde los diseños presentados y los dibuje uno por uno, sin importar el orden en el que los recuerde. Complementariamente, se le solicita resolver ítems de coordinación motora y discriminación visual. La calificación de cada reproducción en "copia" y "recuerdo" se hace mediante el Sistema global de calificación, un método intuitivo y continuo para medir el cambio de la calidad de los diseños reproducidos; el puntaje varía entre 0 (ausencia de forma en el dibujo) y 4 (reproducción casi perfecta). El manual presenta ejemplos de cada categoría de puntuación. Sus aspectos normativos y psicométricos se efectuaron en una muestra de más de 4000 personas estratificadas por etnia, educación y estatus socioeconómico, empatando estas características 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.

Procedimiento

Primero, se procedió a traducir las instrucciones de aplicación y calificación del Bender-II. La traducción al español se hizo por un psicólogo bilingüe, y fue revisado por otros dos psicólogos igualmente fluentes en el idioma inglés; adicionalmente, un psicólogo nativo americano con experiencia en el uso del Bender-II, verificó la traducción. Se juzgó que las traducciones convergieron semánticamente.

El Bender-II se aplicó con un conjunto de otras pruebas de habilidad visomotora y habilidad cognitiva general. La administración de las pruebas se hizo durante un periodo de dos meses y para cada una de ellas, previamente se efectuaron sesiones de entrenamiento consistentes en comprender el uso de las pruebas, practicar las instrucciones de administración y ensayar las normas de calificación. Durante la administración, se siguieron estrictamente las instrucciones del manual, además de procurar satisfacer las condiciones mínimas de administración estandarizada en la interacción individualizada (Lee, Reynolds, & 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. La secuencia de las pruebas del BGT-II se aplicó según lo indicado en el manual: "copia", "recuerdo", "motor" y "percepción". En el presente estudio únicamente se analizarán los puntajes copia y recuerdo.

Respecto a los análisis, la estimación del acuerdo intercalificadores se hizo en el contexto del modelo del análisis de varianza usando el coeficiente de correlación intraclase (ICC) (Shrout & Fleiss, 1979) que se aplica cuando los datos analizados tienen una métrica continua. Se utilizó el modelo 2 para el cálculo de ICC, que asume que los calificadores son seleccionados aleatoriamente de una población de potenciales calificadores y que cada calificador evalúa a cada examinado; este es el modelo de efectos aleatorios de dos vías y cubre mayormente las situaciones de acuerdo intercalificadores (McGraw & Wong, 1996). El reporte de ICC usará la estima-

ción correspondiente para una sola medición (ICC [2,1]). En la determinación del grado de acuerdo, los niveles cualitativos de acuerdo recomendados tienden a variar de autor en autor (Charter, 2003), pero usaremos uno que frecuentemente se encuentra en la literatura que es el de Cicchetti (1994). Este autor declara cuatro niveles de evaluación cualitativa aplicable al acuerdo intercalificadores: < 0.40 = pobre; 0.40 – 0.59 = aceptable; 0.60 – 0.74 = bueno; > 0.74 = excelente.

El análisis de acuerdo tuvo dos fases: la primera para el rango completo de puntajes y la segunda para los grupos extremos de la distribución, que representan los grupos de bajo y alto desempeño. La separación de estos grupos se hizo con la desviación estándar como criterio, es decir, una desviación estándar debajo y sobre la media para identificar los grupos bajo y alto, respectivamente. Sin embargo, esta partición produciría una natural restricción del rango en la dispersión de los puntajes y una consecuente disminución espuria de los coeficientes (Chen & Popovich, 2002). Para atenuar este efecto, se aplicó una corrección por restricción del rango basado en McNemar (1949).[1] El efecto de esta corrección sería la de aumentar la magnitud del coeficiente de acuerdo al grado de restricción de la variabilidad ocurrida en la desviación estándar.

Resultados

Acuerdo en el puntaje total. El acuerdo entre las tres calificadoras arrojó coeficientes mayores a 0.77 en los grupos de cuatro y cinco años y en los puntajes de "copia" y "recuerdo" (tabla 1). El acuerdo fue ligeramente menor en "copia" para el grupo de cinco años, pero no fue estadísticamente significativo. En general, el nivel de acuerdo considerando todo el rango de puntajes mostró un excelente nivel de acuerdo. Las correlaciones Pearson entre cada par posible de calificadores varió entre 0.80 y 0.92 en los grupos de cuatro y cinco años y en los puntajes de "copia" y "recuerdo".

Acuerdo en los puntajes extremos. Al separar la muestra total en dos grupos de rendimiento (alto y bajo), se observaron grandes diferencias en la concordancia de los calificadores. En general, y como se esperaba, los ICC fueron de menor magnitud en los grupos alto y bajo; luego de la corrección por restricción del rango, ocurrieron cambios importantes. Las estimaciones del acuerdo para el puntaje de "recuerdo" fueron consistentemente elevadas en ambos grupos de rendimiento en la edad de cinco años, pero en el puntaje "copia" hubo en efecto diferencial. Esta inconsistencia también ocurrió en los puntajes de "copia" para el grupo de cuatro años en que se obtuvo un excelente acuerdo en los puntajes de grupo de bajo rendimiento (-1DE). El ajuste por restricción del rango produjo un esperado cambio en los coeficientes, pues aumentaron su magnitud en alrededor de 1.36 ($ICC_{corregida}/ICC_{no\text{-}corregida}$). El patrón de acuerdo, sin embargo, fue el mismo que en los coeficientes no corregidos.

Discusión

El presente estudio exploró el acuerdo intercalificadores en los puntajes obtenidos de una muestra de niños preescolares sobre el Bender-II. Este renovado instrumento apenas ha tenido estudios psicométricos y aplicados y no se ha reportado su uso con participantes hispanos en sus propios países. En contraste con lo reportado en su manual y algunos estudios recientes (Volker et al., 2010; Merino, 2012), no sólo se analizó el acuerdo en rango completo de puntajes, sino también en los extremos de la distribución, en donde los puntajes son críticos para el diagnóstico. El acuerdo en el puntaje total de "copia" fue elevado, pero en las regiones extremas tuvo un patrón inestable e interaccionó con la

1 En la siguiente ecuación aplicada (McNemar, 1949), DE_X es la desviación estándar del grupo extremo, DE_x la desviación estándar del grupo total, r_r la correlación restringida y r_{nr} la correlación corregida o no restringida:

$$r_{nr} = \frac{r_r\left(\frac{DE_X}{DE_x}\right)}{\sqrt{1 - r_r^2 + r_r^2\left(\frac{DE_X}{DE_x}\right)^2}}$$

Tabla 1.
Resultados del acuerdo intercalificadores (ICC, modelo 2) en la muestra total y en dos grupos de rendimiento

	Grupo Total	Grupo *+1DE*		Grupo *-1DE*	
		No ajustado	Ajustado[a]	No ajustado	Ajustado[a]
4 años					
Copia	0.86	0.15	0.37	0.81	0.94
5 años					
Copia	0.78	0.45	0.62	0.10	0.33
Recuerdo	0.88	0.43	0.70	0.46	0.98

[a] ajuste McNemar (1949) por restricción del rango. DE: desviación estándar.

Fuente: elaboración propia.

edad. En la edad de cuatro años, el acuerdo fue más difícil de lograr en los desempeños más elevados, mientras que en la edad de cinco años ocurrió en los desempeños más bajos. Estos resultados sucedieron aún luego de efectuarse la desatenuación por restricción por rango de los ICC en las muestras extremas (Zimmerman & Williams, 2000). En esta situación, la corrección de McNemar pudo proporcionar una mejor estimación de este parámetro.

Por otro lado, aunque el pequeño tamaño muestral en las regiones extremas del puntaje elevó el error estándar de los coeficientes, haciéndolos imprecisos, los resultados revelan una importante precaución sobre la interpretación de los puntajes en los grupos en que usualmente se toman decisiones de naturaleza clínica. Usando el Sistema de calificación global del Bender-II, el desempeño muy bajo o muy alto en el puntaje total se asocia a reproducciones en los ítems que pueden ser puntuables entre 0 o 1 y 4 o 5 respectivamente; es posible que estos niveles de desempeño demanden más esfuerzo en la observación clínica de evaluador y que requiera diferenciar más la puntuación más apropiada.

Algunos de los resultados hallados mostraron legítimamente un nivel de acuerdo elevado y que no puede ser explicado por artefactos como la presencia de puntajes raros (*outliers*) que pueden ocasionalmente elevar las estimaciones de acuerdo (Zimmerman & Williams, 2000). El efecto de la edad de los examinados parece interaccionar con la dificultad en lograr un acuerdo, pero independientemente de esta situación, hay consecuencias prácticas importantes para la evaluación clínica que se derivan de nuestros resultados. Por ejemplo, se requeriría un cuidadoso entrenamiento focalizado en mejorar la observación y decisión de puntuación en los niveles de pobre o alto rendimiento visomotor, así como tomar en cuenta la experiencia de los examinadores en el uso de pruebas. Otra alternativa más demandante sería promediar las calificaciones obtenidas en estos grupos de puntajes extremos; pero esto exige más de un calificador si se quiere tener la máxima seguridad en el puntaje obtenido. Tener más calificadores es un costo que el usuario debe balancear entre la obtención de un puntaje cercanamente preciso y confiable, y el tiempo, el entrenamiento y disponibilidad de otros calificadores.

Actualmente se podría orientar la investigación de la confiabilidad del Bender-II hacia la comparación con otros métodos. Las comparaciones entre varios sistemas de calificación para el BGT han demostrado que pueden ser similarmente confiables en diferentes muestras de participantes (Field, Bolton, & Dana, 1982; Lacks & Newport, 1980; McIntosh, Belter, Saylor, Finch, & Edwards, 1988); sin embargo, estos estudios se han enfocado en el rango total de puntuación y con sistemas de puntuación que hoy pueden considerarse antiguas. En el habla hispana, una búsqueda informal de las investigaciones actuales sobre el TGB podría hallar apenas unos cuantos estudios en que la evaluación psicométrica sea el objetivo principal de estudio. Ya que los cambios estructurales y funcionales del Bender-II no lo hacen un "vino antiguo en botella nueva", sino una herramienta muy diferenciada de sus antecesores,

hay una nueva vertiente de investigación que puede ser enfocada principalmente hacia estudios normativos y hacia su posible sensibilidad a las potenciales diferencias interculturales.

Considerando que luego de ocho años de su publicación aún no se han presentado resultados en participantes de otros países, hay un claro horizonte para evaluar la utilidad práctica y teórica de esta nueva versión. Y dado que el nuevo Bender-II demanda más juicio en la aplicación de su sistema de calificación, la evaluación de la variabilidad en la puntuación proveniente de los calificadores es un aporte necesario para interpretar los límites de la validez de los puntajes (Feldt & Brennan, 1989) y evaluar las consecuencias del uso del Bender tal como se establece dentro del marco de la validez consecuencial de las pruebas (Messick, 1995).

Referencias

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME) (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Archer, R. P., & Newsom, C. R. (2000). Psychological test usage with adolescent clients: Survey update. *Assessment, 7*(3), 227-235.

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.* Research Monographs, No. 3. New York: American Orthopsychiatric Association.

Bender, L. (1946). *Instructions for the use of the Visual Motor Gestalt Test*. New York: American Orthopsychiatric Association.

Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet, 8*, 307-310.

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

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.

Brannigan, G. G., & Decker, S. L. (2003). *Bender Visual-Motor Gestalt Test, Second Edition.* Itasca, IL: Riverside Publishing.

Campbell, J. M., Brown, R. T., Cavanagh, S. E., Vess, S. F., & Segall, M. J. (2008) Evidence-based Assessment of Cognitive Functioning in Pediatric Psychology. *Journal of Pediatric Psychology*. Accesado: 13 de Mayo del 2009. Recuperado de http://jpepsy.oxfordjournals.org/cgi/content/full/jsm138v1#B8.

Charter, R. A. (2003). A breakdown of reliability coefficients by test and reliability method, and the clinical implications of low reliability. *Journal of General Psychology, 130*(3), 209-304.

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.

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.

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

Feldt, L. S., & Brennan, R. L. (1989). *Reliability*. In R. H. Linn (Ed.), *Educational Measurement* (3rd Ed.). American Counsel of Education. New York: Macmillan.

Field, K., Bolton, B., & Dana, R. H. (1982). An evaluation of three Bender-Gestalt scoring systems as indicators of psychopathology. *Journal of Clinical Psychology*, *38*(4), 838-842.

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), 468-469.

Hutt, M. L. (1969). *The Hutt Adaptation of the Bender-Gestalt Test* (2nd Ed.). New York: Grune & Stratton.

Hutt, M. L. (1985). *The Hutt Adaptation of the Bender-Gestalt Test* (4th Ed.). New York: Grune & Stratton.

Koppitz, E. M. (1963). The Bender-Gestalt Test for young children (2nd Ed.) New York: Grune & Stratton.

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

Lacks, P. (1999). *Bender-Gestalt screening for brain dysfunction* (2nd Ed.). New York: Wiley.

Lacks, P. B, & Newport, K. (1980). A comparison of scoring systems and level of scorer experience on the Bender-Gestalt Test. *Journal of Personality Assessment*, *44*(4), 351–357.

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.

McIntosh, J. A., Belter, R. W., Saylor, C. F., Finch, A. J., Jr., & Edwards, G. L. (1988). The Bender-Gestalt with adolescents: Comparison of two scoring systems. *Journal of Clinical Psychology, 44*(2), 226-30.

McNemar, Q. (1949). *Psychological statistics*. New York: Wiley.

Merino, C. (2009). Un análisis no paramétrico de ítems de la prueba del Bender Modificado para estudiantes de primaria. *Liberabit, 15*(2), 83-94.

Merino, C. (2010). El sistema de calificación cualitativa para la Prueba Gestáltica de Bender – Modificada: Estudio preliminar de sus propiedades psicométricas. *Avances en Psicología Latinoamericana, 28*(1), 63-73.

Merino, C. (2011). Construct validity of the Qualitative Grading System for the Modified Bender Gestalt Test. *Electronic Journal of Research in Educational Psychology*, 9(3), 1245-1266.

Merino, C. (2012). Fiabilidad en el Test Gestáltico de Bender – 2da versión, en una muestra independiente de calificadores. *Revista de Investigación Educativa*, *30*(1), 222-232.

Merino, C., & Benites, L. (2011). Evaluación de la confiabilidad del Sistema Cualitativo de Calificación para la versión modificada del Test Gestáltico de Bender. *Universitas Psychologica*, *10*(1), 231-243.

Messick, S. (1995). Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741-749.

Nunnally, J.C. & Bernstein, I.J. (1995). *Teoría Psicométrica*. México, D.F.: McGraw-Hill Interamericana.

Parsons, L., & Weinberg, S. L. (1993). The Sugar Scoring System for the Bender Gestalt Test: An objective approach that reflects clinical judgment. *Perceptual and Motor Skills, 77,* 883-893.

Pauker, J. D. (1976). A quick-scoring system for the Bender-Gestalt: Interrater reliability and scoring validity. *Journal of Clinical Psychology, 32*, 86-89.

Perticone, E. X. (1998). *The Clinical and Projective Use of the* Bender–*Gestalt Test*. Springfield, IL: Charles Thomas.

Piotrowski, C. (1995). A review of the clinical and research use of the Bender-Gestalt Test. *Perceptual and Motor Skills, 81*, 1272-1274.

Raphael, A. J., & Reichenberg, N. (1992). *Advanced psychodiagnostic interpretation of the Bender Gestalt Test: Adults and children*. Connecticut: Praeger Publishers.

Reynolds, C. R. (2007). *Koppitz Developmental Scoring System for the Bender Geslalt Test (KOPPITZ-2).* Austin, TX: Pro-Ed

Salvia, J., Ysseldike, J., & Bolt, S. (2009). *Assessment: In Special and Inclusive Education* (11th Ed.). Belmonth, CA: Wadsworth.

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*. São Paulo: Vetor Editora.

Sullivan, K., & Bowden, S. C. (1997). Which tests do neuropsychologist use? *Journal of Clinical Psychology, 53,* 657-661.

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.

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, C. E., Jr., Campbell, V. L., Nieberding, R., & Hallmark, R. (1995). Contemporary practice of psychological assessment by clinical psychologists. *Professional Psychology: Research and Practice, 26,* 54-60.

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

Zimmerman, D. W., & Williams, R. H. (2000). Restriction of range and correlation in outlier-prone distributions. *Applied Psychological Measurement*, *24*(3), 267-280.