Interdisciplinaria
ISSN: 0325-8203
interdisciplinaria@fibercorp.com.ar
Centro Interamericano de Investigaciones Psicológicas y Ciencias Afines
Argentina

MERINO SOTO, CÉSAR; ALLEN, RYAN A.
CONFIABILIDAD INTERCALIFICADORES Y VALIDEZ DE CONSTRUCTO DEL TEST GESTÁLTICO DE BENDER (SEGUNDA VERSIÓN)
Interdisciplinaria, vol. 30, núm. 2, 2013, pp. 253-264
Centro Interamericano de Investigaciones Psicológicas y Ciencias Afines
Buenos Aires, Argentina

Confiabilidad intercalificadores y validez de constructo del Test Gestáltico de Bender (segunda versión)*

Inter-scorer reliability and construct validity in the Bender Gestalt Test (second version)

César Merino Soto** y Ryan A. Allen***

*La investigación que se informa fue respaldada por el Instituto de Investigación de Psicología de la Universidad de San Martín de Porres (Lima, Perú).
**MA en Psicología. Docente e investigador en el Instituto de Investigación de Psicología de la Universidad de San Martín de Porres. E-Mail: sikayax@yahoo.com.ar
Instituto de Investigación de la Universidad de San Martín de Porres. Av. Tomás Marsano 242, 5º Piso. Lima 34, Perú.
**PhD. Profesor Asistente en el Department of Education & Allied Studies de la John Carroll University.
E-Mail: rallen@jcu.edu
Department of Education & Allied Studies, John Carroll University, University Heights, OH 44118. United States.
Los autores agradecen la colaboración recibida a los participantes del estudio y a las autoridades de la institución educativa que permitió la aplicación del instrumento.

Resumen

Hace algunos años, se ha publicado el nuevo Test Gestáltico de Bender, segunda versión (*Bender-II*). Esta nueva versión usa el Sistema de Calificación Global (SCG) para obtener los puntajes, que enfatiza la reproducción exacta de los diseños; junto con otros cambios estructurales, consiste en la mayor modificación realizada al Test de Bender para evaluar la habilidad visomotora en varias etapas del desarrollo humano. Pero aún hay pocas investigaciones que han estudiado su funcionamiento psicométrico en muestras hispanas y especialmente el error de medición en el procedimiento de la calificación.

Se investigó el efecto de la variabilidad de la calificación del nuevo Bender-II sobre su *validez* de constructo con una medida de inteligencia. Se administraron el Bender-II y el Test Breve de Inteligencia de Kaufman (*K-BIT*) a 60 niños preescolares, y tres calificadores calificaron las figuras reproducidas mediante procedimientos estandarizados y el SCG. El análisis comparó las correlaciones obtenidas por cada calificador entre los puntajes del Bender-II y del K-BIT. Se halló que la variabilidad de los puntajes proveniente de la interpretación del SCG afectó moderadamente las correlaciones con el puntaje total y de las subescalas del K-BIT. Un calificador fue menos confiable y sus puntajes mostraron al menos 5% menos variancia compartida al compararlo con los otros calificadores. Se concluye que la interpretación de la validez se puede distorsionar aun cuando el error de medición es moderado pues interaccionan con otras fuentes de error. Esto sugiere que se debe garantizar un buen acuerdo en la calificación de pruebas que requieran juicio.

Palabras clave: Validez; Visomotricidad; Bender-II; Confiabilidad; K-BIT.

Abstract

Although the impact of the measurement error in the accuracy of the classification of subjects and validity correlations is theoretically established, in a practical situation has not been explored the degree of impact on the new *Bender-II* (Brannigan & Decker, 2003). The practical situation of assessment is that which occurs in the professional context, in which one or more examiners of an assessment team assessing children in a particular institution.

The new Bender-II has substantial modifications in its structure and functioning, making it different from the original instrument proposed by Bender (1938, 1946). Structural changes consisted in more of items (16 designs), complementary tests (fine motor and visual perception) and two major tests (Visual Constructive Memory and *Visual Motor*), the rating method Global Scoring System (GSS) and standardized record sheet for the child's behavior during the test administration. The GSS was created ad hoc for the Bender-II, and it is a method that emphasizes the exact reproduction of the designs; its origin is in the original gestalt approach of Bender. Main studies have been published in the manual (Brannigan & Decker, 2003), and subsequent studies have used the American standardization sample. However, in non-immigrant Hispanic population, to date there are some unpublished and published only one (Merino, 2012); therefore, it is not known how generalizable the findings and psychometric properties obtained in the American standardization sample are.

The aim of the study was to examine the effect of the variability of the scoring of designs of the new Bender-II on construct *validity* with a measure of intelligence. The Bender-II and Kaufman Brief Intelligence Test (*K-BIT* - Kaufman, A.S. & Kaufman, A.L., 1994) was administered to 60 preschool children (between 4 and 5 years, 33 girls), and three scorers rated the designs reproduced, by standardized procedures and GSS. The analysis consisted of two steps: first, we estimated the consistency and inter-rater agreement using random two-way intraclass correlations (McGraw & Wong, 1996; Shrout & Fleiss, 1979). Second, for each scorer, correlations were calculated between scores on the Bender-II and K-BIT, and finally these correlations were compared with a test for dependent correlations with a common element (Steiger, 1980, Williams, 1959).

The results indicate that there were slight differences between two scorers, but one of them had comparatively lower coefficients of consistency and agreement. In the all scorers, the mag nitude of the consistency coefficients (> .85) and agreement (> .84) between the qualifiers indicate good levels of concordance with even moderate exercise time (two or three sessions). Correlations between scores on the Bender-II and K-BIT were around .43 and .61, and the lowest correlations occurred in the scorer that showed less consistency and agreement with other scorers, clearly indicating the impact of measurement error in the validity correlations. Compared with one of the rating, this difference was statistically significant, and the percentage reduction of covariance was at least 5%.

Finally, the results indicate several points. First, we present other evidence for inter-rater *reliability* of the Bender-II, and that there were good levels of agreement and consistency. Second, there is potentially a reduction in the correlations of validity when a scorer has trouble for interpret consistently and correctly Global Scoring System. Correlations decreased, however, are not equal in all the scorers, and therefore must be verified the goodness of fit between scorer and the Bender it's qualification method approach. The results can be are idiosyncratic to the sample and study conditions, and the sample size constraints threaten the generalizability of the findings. However, the study conditions are close to professional practice and therefore can be generalized to some extent. In addition, can serve as a baseline to compare future studies of reliability in the Bender-II.

Key words: Validity; Visual motor; Bender-II; Reliability; K-BIT.

Desde la creación del Test Gestáltico Visomotor de Bender (TGB - Bender, 1938, 1946) se han propuesto diferentes sistemas de calificación y cambios estructurales para intentar mejorar su función diagnóstica (Merino, 2011a), lo cual ha motivado que di-

versos investigadores realicen estudios sobre su validez y confiabilidad. El TGB inicialmente contaba con nueve láminas, las cuales se calificaban con variados sistemas; uno de los más populares para niños entre 5 y 10 años fue el Sistema Evolutivo de Koppitz (1984), que puntuaba el error en los diseños reproducidos con un intervalo de 0 a 1. En 2003, Brannigan y Decker hicieron una extensa revisión de la prueba y adicionaron siete láminas a la segunda versión del TGB (Bender-II), además de nuevos procedimientos para explorar la motricidad, percepción visual, memoria visual y comportamiento durante la evaluación.

El Bender-II (B-II) se puede aplicar a personas de 4 a 80 años y actualmente es la generación siguiente del tradicional TGB publicada hace más de 70 años por Bender (1938, 1946), cuya popularidad entre los psicólogos se extendió hasta la década pasada (Watkins, Campbell, Nieberding & Hallmark, 1995). El B-II tiene la ventaja de ser un instrumento único para un amplio rango de edad, con aspectos estructurales y funcionales también únicos que extienden su funcionalidad para la evaluación neuropsicológica. Por ejemplo, actualmente contiene cuatro fases de evaluación: la primera es el copiado, luego, el recuerdo, la producción motora y la percepción visual. Estos últimos componentes estructurales amplían los usos del TGB dentro de una evaluación integrada para la detección de problemas de memoria, visomotricidad y la influencia recibida de la motricidad y percepción visual, además de indicadores objetivos del tiempo de respuesta y comportamiento durante la evaluación.

El copiado de las figuras geométricas es una de las tareas más frecuentes para evaluar la visomotricidad (Cummings, Hoida, Machek & Nelson, 2003), pero su calificación requiere un monto de subjetividad que depende del aprendizaje del sistema de calificación y de su validez de contenido. Esto es cierto para el TGB ya que la variabilidad de sus puntajes es influenciada por la variabilidad de los calificadores en la aplicación del sistema de calificación. Esta fuente de potencial error es una de las más relevantes para instrumentos que requieren la subjetividad del examinador para determinar el nivel de desempeño evaluado (Feldt & Brennan, 1989). El Bender-II incluye un nuevo sistema para calificar las reproducciones, es el Sistema de Calificación Global (SCG), que puede requerir un alto grado de discernimiento por parte del examinador para determinar la puntuación más apropiada a cada diseño. Este sistema es inédito y propio del B-II y se fundamenta en el trabajo original de Bender (1938) con respecto a la inspección holística de los diseños. Su aplicación se enfoca en la determinación de la similaridad de la reproducción ejecutada por el examinado y el estímulo presentado, requiriendo una perspectiva holística para identificar el nivel de puntuación. Los estudios de validez y confiabilidad publicados en el manual del B-II (Brannigan & Decker, 2003) indican que sus propiedades psicométricas en población americana pueden considerarse buenas. Sin embargo, apenas se han realizado investigaciones publicadas y no publicadas que repliquen estos resultados en población hispana (Merino, 2011a, 2011b, 2011c). Entre los últimos estudios realizados, en Puerto Rico Cruz (2008) obtuvo normas preliminares y en Perú, Manzanares y Merino (2013) verificaron la relación no lineal entre edad y visomotricidad y Merino (2011b) reportó un análisis de ítemes en una muestra independiente, también en Perú.

De los estudios hispanos, solo uno publicado se enfocó en el acuerdo de intercalificadores (Merino, 2012) y obtuvo una buena confiabilidad para el puntaje visomotor y confiabilidad moderada para los ítemes. Por otro lado, los estudios empíricos anglosajones con el Bender-II obtuvieron descripciones inéditas del desarrollo visomotor desde los 4 a los 80 años (Decker, 2007), asociaciones multivariadas con la inteligencia verbal y no verbal (Decker, Allen & Choca, 2006), comparaciones concurrentes con el VMI-5 (Volker et al., 2010) y los correlatos evolutivos y cognitivos (Decker, Englund, Carboni & Brooks, 2011). Gran parte de las investigaciones anglosajonas (Decker, 2007; Decker, Allen, & Choca, 2006; Decker et al.,

2011) se basaron en la muestra de estandarización obtenida para el manual.

Como parte de las condiciones para establecer la garantía métrica de un instrumento de medición, los estudios de confiabilidad son importantes porque el error es inherente a todo proceso de medición (Feldt & Brennan, 1989) y su estimación puede informar sobre su impacto y sobre las estimaciones de validez. La validez estimada por un coeficiente de correlación tiende a disminuir mientras mayor es el error de medición (Feldt & Brennan, 1989). Considerando el uso del SCG del Bender-II, la variabilidad en la aplicación de tal sistema por parte de diferentes calificadores es una fuente de error relevante, pues ello es típico en instrumentos en los que se involucra el juicio del examinador para obtener los puntajes. Esta variación en asignar puntajes diferentes a los mismos protocolos puede ser consecuencia de la interpretación inusual de un calificador, de la insuficiente claridad de las instrucciones de calificación, de la dificultad del diseño reproducido, o de una interacción entre todas ellas. En cualquier situación, esta variación introduce error de medición, específicamente sobre las correlaciones de validez para establecer la red nomológica del Bender-II**.**

Objetivos e hipótesis

El objetivo del estudio que se informa fue examinar el efecto de la confiabilidad entre calificadores sobre los coeficientes de validez, originados por la variabilidad en la calificación, enfocándose en el Bender-II (Brannigan & Decker, 2003) y su relación con un instrumento de habilidad intelectual.

Se hipotetiza que: (a) las correlaciones de validez entre el Bender-II y una prueba de habilidad intelectual disminuirán como efecto de la variación entre-calificadores en el uso del SCG y (b) que la variación entre calificadores está asociada con una disminución en la consistencia interna de los puntajes del Bender-II.

Estos objetivos sustanciales se desarrollaron en el contexto de uno de los primeros usos publicados del Bender-II en habla hispana y pretende contribuir acumulativamente a sus evidencias de validez.

Por otro lado, la asociación entre el Bender-II y una medida de habilidad intelectual es relevante por los vínculos teóricos neuropsicológicos, que actualmente han fundamentado la validez de constructo del Bender-II (Decker et al., 2006; Decker et al., 2011).

En el estudio realizado se usó el puntaje visomotor del Bender-II, que es componente principal del test.

Método

Muestra

Participaron dos grupos: uno de niños y otro de calificadores.

Los niños fueron 60 preescolares de ambos sexos (33 niñas, 55%) distribuidos similarmente entre 4 (n = 30, 50%) y 5 años de edad, todos procedentes de un centro educativo público de educación regular, ubicado en una zona urbana de Lima Metropolitana (Perú). Se solicitó la autorización de la institución educativa y el consentimiento de los padres. Los niños no formaron parte de alguna intervención *ad hoc* orientada al mejoramiento de las habilidades visuales o motoras en su casa o en el colegio. Su nivel socioeconómico, estimado por el reporte del director y profesores, era medio y sus padres tenían educación básica completa y en algunos casos, estudios superiores.

Los calificadores fueron tres estudiantes (identificados como A, B y C) de la Facultad de Psicología de una universidad privada, pertenecientes al tercio superior de rendimiento académico. Su colaboración fue voluntaria y en el marco de actividades extracurriculares para el aprendizaje de nuevas pruebas psicológicas.

Instrumentos

Test Gestáltico Visomotor de Bender, 2da. Versión (Brannigan & Decker, 2003)

Es la nueva versión aplicable a sujetos de 4 a 80 años de edad y evalúa el funciona-

miento visomotor y otros aspectos del funcionamiento cognitivo asociados a la visomotricidad tales como la motricidad, la percepción visual y la memoria visual constructiva. Contiene 16 láminas que incluyen nuevos diseños añadidos para extender el escalamiento del puntaje.

El Bender-II contiene dos fases: Copia y Recuerdo y dos pruebas suplementarias que evalúan la motricidad fina y la percepción visual. Puntúa las reproducciones con el Sistema de Calificación Global (SCG), asignando un puntaje entre 0 y 4 para evaluar el grado de exactitud del diseño dibujado. Este instrumento fue normalizado en 4.000 personas estratificadas por etnia, educación y estatus socioeconómico en población norteamericana (Brannigan & Decker, 2003).

Test Breve de Inteligencia de Kauffman (K-BIT, Kaufman, A.S. & Kaufman, A.L., 1994)

Es una prueba breve, de aplicación individual para personas de 4 a 90 años y orientada a la medición de la inteligencia general mediante el subtest verbal (Vocabulario expresivo y Definiciones) y otro no verbal (Matrices), ambos relacionados teóricamente con la inteligencia cristalizada y fluida, respectivamente; dicho modelo sirvió como soporte para su construcción. Aunque el reducido muestreo de las habilidades intelectuales lo limita para ser aplicado como instrumento único en la descripción individual (Chin et al., 2001), estudios independientes sobre su validez concurrente en poblaciones especiales han demostrado la eficiencia diagnóstica de su puntaje CI total (Canivez, 1995) y adecuada validez convergente y divergente con medidas de habilidad cognitiva y problemas de conducta (Canivez, Neitzel & Martin, 2005; Childers, Durham & Wilson, 1994). Su asociación con otras medidas de inteligencia también ha sido alta (Bowers & Pantle, 1998; Grados & Russo-Garcia, 1999; Bowers & Pantle, 1998). El manual reporta que los estudios de validez en la población angloparlante y en la adaptación hispana muestran propiedades psicométricas satisfactorias para despistaje en adultos, adolescentes y niños (Kaufman, A.S. & Kaufman, A.L., 1994).

Una extensa revisión de las investigaciones independientes puede inspeccionarse en Canivez y colaboradores (2005).

Para el estudio realizado, se modificaron algunos ítemes de la Escala de Vocabulario Expresivo de acuerdo a los resultados de Zayerz y Manco (2008), lo que consistió en reemplazar palabras de uso no frecuente en la adaptación española del K-BIT.

Procedimiento

El contexto de la investigación fue la validación del Bender-II en población peruana. Esta prueba se aplicó con un conjunto de otras medidas de habilidad visomotora y habilidad cognitiva general, algunas administradas en forma grupal y otras, individualmente. La administración de las pruebas fue realizada durante un período de dos meses por el autor de este artículo y tres estudiantes, quienes fueron previamente entrenados en tres sesiones para la aplicación y calificación de los instrumentos, monitoreándose el desarrollo de sus evaluaciones y calificaciones.

Durante la administración se siguieron estrictamente las instrucciones de los manuales, manteniendo las condiciones óptimas para la aplicación individual de los instrumentos. Estas consideraciones fueron enfatizadas para minimizar el impacto de potenciales eventos no relevantes durante la evaluación (Bracken, 2007; Lee, Reynolds & Willson, 2003).

Todos los niños participantes fueron evaluados con el Bender-II, pero dos de ellos no participaron de la administración del K-BIT; por lo tanto, las correlaciones de validez solo se calcularon para 58 niños de la muestra. Los análisis involucraron la obtención de correlaciones Pearson entre el Bender-II y K-BIT, la comparación de estas correlaciones y la estimación del acuerdo intercalificadores, cuyos procedimientos serán descriptos en Resultados.

Para la descripción cualitativa del grado de acuerdo, se trabajó con los niveles pro-

puestos por Cicchetti (1994), que son los más frecuentemente utilizados en la literatura para coeficientes de correlación intraclase (≤ .39: pobre, ≤ .59: aceptable, ≤ .74: bueno, y ≥ .75: excelente - Halgren, 2012).

Resultados

Los estadísticos descriptivos se presentan en la Tabla 1 y se observa la similaridad de los puntajes medios y la dispersión en cada calificador usando el SCG. Para examinar la consistencia y el acuerdo entre los calificadores, se usó el modelo de dos vías aleatorias del coeficiente de correlación intraclase (ICC - McGraw & Wong, 1996; Shrout & Fleiss, 1979), que considera a los sujetos y calificadores como una muestra de la población a la cual generalizar los resultados. Este modelo también fue elegido por Merino (2012), considerando que correspondería a la mayoría de las evaluaciones del acuerdo entre calificadores en la investigación (Mc Graw & Wong, 1996). La unidad de análisis fueron las calificaciones individuales y se evaluó específicamente el acuerdo absoluto (ICC_{ac}) y la consistencia (ICC_{con}).

Los índices de acuerdo (ICC_{ac}) y consistencia (ICC_{con}) indicados por dos de los tres calificadores (A y C) muestran valores elevados y satisfactorios, lo cual sugiere que el aprendizaje y el uso del SCG pueden producir puntajes altamente confiables (ver Tabla 2). Sin embargo, se observa que uno de los calificadores (B) fue constantemente bajo en los índices de acuerdo y consistencia calculados. Estos niveles relativamente bajos ocurrieron en los análisis apareados entre este calificador y el resto (A y C) y se dirige a respaldar la segunda hipótesis planteada en esta investigación.

En las correlaciones de validez (ver Tabla 2) entre el Bender-II y los puntajes de K-BIT (K-BIT-Total, K-BIT-E: Expresión y K-BIT-M: Matrices) se observa la variación de sus magnitudes. Las correlaciones del B-II con el K-BIT-E y el K-BIT-Total para el calificador B, fueron relativamente bajas con respecto a los otros calificadores.

Para analizar si las correlaciones fueron estadísticamente diferentes, se aplicó una prueba *t* de comparación de correlaciones dependientes con un elemento común (Steiger, 1980; Williams, 1959). Aquí, el elemento común en las variables de las dos correlaciones comparadas son los puntajes del Bender-II en cada par de calificadores que son comparadas. Considerando la correlación entre K-BIT-E y Bender-II, solo hubo una diferencia estadísticamente significativa entre los calificadores A y B [$t(55) = 2.24$, $p = .029$]. En la relación K-BIT-M y Bender-II no ocurrieron diferencias significativas entre los calificadores; similarmente, en la correlación del puntaje total del K-BIT con Bender-II, la máxima diferencia entre las correlaciones no fue estadísticamente significativa [$t(55) = 1.93$, $p = .059$].

Se observa también que la variancia compartida entre el Bender-II y las pruebas de inteligencia es menor en el calificador B, quien también obtuvo menos acuerdo con el resto de los calificadores. Específicamente, la variancia compartida entre el K-BIT y el Bender-II en el calificador B es en promedio, 5% menos que las otras correlaciones obtenidas en los otros calificadores.

Discusión

Se presenta un avance de la investigación realizada sobre la nueva versión del Bender. Los estudios hispanos aún no prestaron atención a esta prueba que ha dado un cambio definitivo a su estructura y funcionalidad.

El objetivo del estudio que se informa fue examinar la variabilidad de la calificación en el nuevo Bender-II y su efecto sobre las correlaciones de validez de constructo. Los resultados pueden separarse en tres aspectos: la confiabilidad / acuerdo, el impacto de este último sobre los coeficientes de validez y la relación entre la habilidad intelectual y visomotricidad.

Sobre el primer punto, la consistencia y el acuerdo son aspectos complementarios y cada uno da una respuesta diferente para determinar la confiabilidad entre-calificadores

(McGraw & Wong, 1996). En el estudio realizado, los resultados de estos aspectos de la confiabilidad entre-calificadores fueron satisfactorios y sugieren que el acuerdo logrado entre usuarios del SCG puede llegar a mostrar una alta similitud entre sus puntajes. Esta similtud establece que el uso del sistema para asignar unívocamente los puntajes llega a buenos niveles de acuerdo y confiabilidad. Debe señalarse que este acuerdo tiene como antecedente una preparación de varias horas y que los calificadores fueron estudiantes sin conocimiento previo de instrumentos de visomotricidad. El profesional aplicado puede no tener una experiencia similar de entrenamiento, aunque sí la experiencia clínica, por lo tanto, la generalización de los resultados debe considerar estos puntos.

En segundo lugar, se observó que la variabilidad entre-calificadores tuvo un ligero impacto sobre los coeficientes de validez entre el Bender-II y el K-BIT. Los puntajes que mostraron más bajo acuerdo en la calificación se asociaron a correlaciones de validez relativamente bajas. La mayoría de las diferencias en la validez no fueron estadísticamente significativas, sugiriendo que estas pueden ser consideradas como variaciones de muestreo y no ser replicables. Sin embargo, el investigador preocupado en la precisión debe considerar que a medida que el error en la calificación aumenta, disminuye la estimación de otros aspectos, como las correlaciones. Esta precisión puede ser importante para acumular meta-analíticamente las evidencias de validez del nuevo Bender-II. También se debe tomar en cuenta el poder estadístico del estudio, porque con un mayor tamaño muestral aumenta la sensibilidad para detectar ensayos estadísticamente significativos.

En tercer lugar, nuestros resultados contribuyen a la validez de constructo del Bender-II al explorar la relación entre habilidades visomotrices y habilidades cognitivas. Las correlaciones obtenidas corroboran los hallazgos multivariados de la actual relación entre ambas (Decker et al., 2006; Decker et al., 2011). Las correlaciones halladas en el estudio fueron altas para los puntajes totales y señalan que la variancia común entre ellas puede tener relevancia diagnóstica y descriptiva en las edades muestreadas. Efectivamente, esto sugiere que el involucramiento de las habilidades visomotrices sobre el desempeño intelectual tiene un fuerte componente lineal y positivo y la interinfluencia de ambos en la resolución de tareas no verbales sería notoria. Sin embargo, en comparación con el puntaje de Matrices, es interesante destacar que se hallaron correlaciones ligeramente mayores entre los puntajes del Bender-II y de Expresión del K-BIT; esto no se corresponde teóricamente con los estudios que han demostrado que el Bender-II tiene mayores asociaciones con las subescalas de habilidades fluidas del WISC-III (Decker et al., 2006) y del Stanford-Binet-V (Decker et al., 2011), correlaciones menores frente a los puntajes de habilidades cristalizadas. Es posible que las habilidades visomotoras de los niños muestreados estén asociadas a prácticas educativas que enfatizan la resolución de problemas, la comprensión de materiales verbales y una mayor oportunidad para el desarrollo del lenguaje (Iverson, 2010), de tal modo que ambos constructos compartan mayor variancia común. Sin embargo, para explicar esta variación se requiere una exploración en una muestra más grande y heterogénea con respecto a la edad.

La creación del Bender-II pone un punto de diferenciación importante entre toda la producción científica y profesional anterior a su publicación, y los aún escasos hallazgos contemporáneos con esta nueva versión sugieren que es una herramienta de potencial interés para la práctica profesional y de investigación. Recientes investigaciones psicométricas en habla hispana (Merino, 2011a, 2012) indican que las propiedades psicométricas con respecto a la confiabilidad por consistencia interna y acuerdo intercalificadores del Bender-II son adecuadas en niños, pero aún son escasas las evidencias de validez en muestras independientes. Aunque tradicionalmente ha sido utilizada como una medida importante para detectar el daño cerebral (Rhodes, D'Amato & Rothlisberg,

2009), se requieren evidencias para el Bender-II sobre las inferencias diagnósticas que pueden hacerse y sobre su eficacia dentro de una batería de evaluación.

El examinador debe tomar en cuenta que existe la potencial variabilidad en la interpretación de sistemas de calificación del TGB y que influye en la precisión de resultados para fines teóricos y aplicados, y por lo tanto, debería estimar el grado de confiabilidad y preparar modos para atenuarlo durante la aplicación y entrenamiento del TGB. Otros métodos pueden incluir más de cuatro calificadores (Svensson & Hill, 1990) o aplicar métodos más sofisticados como la teoría de la generalizabilidad (Rae & Hyland, 2001) para descomponer la fuente de variación en los puntajes.

Finalmente, cabe mencionar las limitaciones del presente estudio. En primer lugar, los calificadores fueron estudiantes de pregrado (como ocurrió en otros estudios recientes, por ejemplo, Merino, 2012) y el resultado del acuerdo no podría generalizarse a usuarios con estatus más alto (docentes, profesionales, investigadores) que ejercerían todas las funciones profesionales del psicólogo; sin embargo, los niveles de acuerdo alcanzados en este estudio son similares a los obtenidos en investigaciones previas con el Bender-II (Brannigan & Decker, 2003; Merino, 2012) y demuestran la eficacia del Sistema de Calificación Global del Bender-II para facilitar altos niveles de confiabilidad. En segundo lugar, no se controlaron los efectos de la edad sobre la asociación entre el K-BIT y el Bender-II, pues una parte de la variancia común entre ambos puede ser explicada por la edad. Debido a la restricción del rango de edad de la muestra, no se podría detectar suficiente variabilidad estadística, afectando a las correlaciones; por tal motivo, se requeriría replicar los resultados en una muestra más heterogénea con respecto a la edad. Y en tercer lugar, la estabilidad de los resultados está comprometida hasta que no se repliquen las relaciones investigadas en otros grupos de niños.

TABLA 1
ESTADÍSTICOS DESCRIPTIVOS DE LOS PUNTAJES ANALIZADOS

	n	Min.	Máx.	*M*	*DE*
Bender-II					
Calificador A	60	0	32	17.02	7.16
Calificador B	60	0	36	16.13	6.99
Calificador C	60	0	36	16.47	7.93
K-BIT					
K-BIT - Expresión	58	6	31	18.07	5.68
K-BIT - Matrices	58	7	23	13.76	3.86
K-BIT - Total	58	16	49	31.83	8.17

TABLA 2
CONSISTENCIA INTERNA Y ACUERDO ENTRE CALIFICADORES (A, B Y C) Y COEFICIENTES DE VALIDEZ BENDER-II (PUNTAJE VISOMOTOR) Y PUNTAJES EN EL K-BIT

	Bender-II A	Bender-II B	Bender-II C
Acuerdo (ICC_{ac})			
Bender-II A	1.0		
Bender-II B	.89**	1.0	
Bender-II C	.92**	.85**	1.0
Consistencia (ICC_{con})			
Bender-II A	1.0		
Bender-II B	.90**	1.0	
Bender-II C	.93**	.86**	1.0
Correlaciones de validez K-BIT – B-II			
K-BIT- Expresión	.54**	.43**	.53**
K-BIT-Matrices	.46**	.48**	.45**
K-BIT-Total	.61**	.52**	.58**

Nota:
ICC_{con}: estimación de la consistencia
ICC_{ac}: estimación del acuerdo
**$p < .01$

Referencias bibliográficas

Bender, L. (1938). A visual-motor gestalt test and its clinical use. *American Orthopsychiatric Association Research Monographs, 3.*

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

Bowers, T.L. & Pantle, M.L. (1998). Shipley Institute for Living Scale and the Kaufman Brief Intelligence Test as screening instruments for intelligence. *Assessment, 5*(2), 187-195. http://dx.doi.org/10.1177/107319119800500209

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

Canivez, G.L. (1995). Validity of the Kaufman Brief Intelligence Test Comparisons with the Wechsler Intelligence Scale for Children - Third Edition. *Assessment, 2*(2), 101-111. http://dx.doi.org/10.177/107319119500200201

Canivez, G.L. Neitzel, R. & Martin, B.E. (2005). Construct validity of Kaufman Brief Intelligence Test, Eschler Intelligence Scale for Children - Third Edition, and Adjustment Scales for Children and Adolescents. *Journal of Psychoeducational Assessment, 23,* 15-34. http://dx.doi.org/10.1177/0734282905023001 02

Childers, J.S. & Durham, T.W. (1994). Relation of performance on the Kaufman Brief Intelligence Test with the Peabody Picture Vocabulary Test-Revised among preschool children. *Perceptual and Motor Skills, 79*(3), 1195-1199. http://dx.doi.org/10.2466/pms.1994.79.3.1195

Chin, C.E., Ledesma, H.M.L., Cirino, P.T., Sevcik, R.A., Morris, R.D., Frijters, J.C. & Lovett, M.W. (2001). Relation between Kaufman Brief Intelligence Test and WISC-III Scores of children with RD. *Journal Learning of Disabilities, 34*(1), 2-8. http://dx.doi.org/10.1177/0022219 40103400101

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

Cruz, D. (2008). *Desarrollo de normas para la Prueba de Desarrollo Viso - Motor Bender II en estudiantes puertorriqueños de 12, 13 y 14 años* [Development of norm for Bender Gestalt Test II in Puerto Rican 12, 13 and 14 years old students]. Trabajo presentado en el Congreso de Medición: Innovación, tecnología y nuevas prácticas en la psicometría. Asociacion de Psicología de Puerto Rico, Universidad Central de Bayamon.

Cummings, J.A., Hoida, J.A., Machek, G.R. & Nelson, J.M. (2003). Visual-motor assessment of children. En C.R. Reynolds & R.W. Kamphaus (Eds.), *Handbook of psychological and educational assessment of children: Intelligence, aptitude, and achievement* (2da. ed., pp. 611-653). 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. http://dx.doi.org/10.1177/07342829 07300685

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.

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

Grados, J.J. & Russo-Garcia, K.A. (1999). Comparison of the Kaufman Brief Intelligence Test and the Wechsler Intelligence Scale for Children - Third Edition in economically disadvantaged African American youth. *Journal of Clinical Psychology, 55*(9), 1063-1071. htttp://dx.doi.org/10.1002/(SICI)1097-4679(19990 9)55: 9<1063::AID-JCLP4>3.0.CO;2-U

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

Iverson, J.M. (2010). Developing language in a developing body: The relationship between motor development and language development. *Journal of Child Language, 37*(2), 229-261. http://dx.doi.org/10.1017/S03050009099 90432

Kaufman, A.S. & Kaufman, A.L. (1994). *K-BIT: Test Breve de Inteligencia de Kaufman. Manual de interpretación* [K-BIT: Kaufman Brief Intelligence Test]. Madrid: TEA.

Koppitz, E.M. (1984). *El Test Gestáltico de Bender: Evolución y nuevas versiones* [The Bender Gestalt Test] (10a. ed.). Buenos Aires: Guadalupe.

Lee, D., Reynolds, C.R. & Willson, V.L. (2003). Standardized test administration: Why bother? *Journal of Forensic Neuropsychology, 3,* 55-81. http://dx.doi.org/10.1300/J151v03n03_04

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.1.1.30

Merino, C. (2010). El sistema de calificación cualitativa para la Prueba Gestáltica de Bender - Modificada: Estudio preliminar de sus propiedades psicométricas [Qualitative Scoring System for the Bender Gestalt Test - Modified: Preliminar study of psychometric properties]. *Avances en Psicología Latinoamericana, 28* (1), 63-73.

Merino, C. (2011a). *El Test Gestáltico de Bender: Evolución y nuevas versiones* [Bender Gestalt Test: Evolution and new versiones]. Trabajo presentado en el III Congreso Internacional de Psicología. Universidad Autónoma del Perú. Lima, Perú.

Merino, C. (2011b). *Avances en la adaptación de la 2da versión del Test Gestáltico Visomotor de Bender* [Advances in the adaptation of Bender Gestalt Test, 2nd version]. Trabajo presentado en el XV Congreso Nacional de Psicología. Lima, Perú.

Merino, C. (2011c). Exploración de diferencias normativas en el Sistema de Calificación Cualitativa para el Test Gestáltico de Bender Modificado [Exploration of normative differences in the Qualitative Scoring System for the Bender Gestalt Test Modified]. *Liberabit, 17*(2), 199-209.

Merino, C. (2012). Confiabilidad en el Test Gestáltico de Bender - 2da versión, en una muestra independiente de calificadores [Reliability in the Bender Gestalt Test, 2nd versión, in a independent sample of scorer]. *Revista de Investigación Educativa, 30*(1), 223-234. http://dx.doi.org/10.5027/psicoperspectivas-Vol12-Issue 1-fulltext-201

Merino, C. & Manzanares, E. (2013, Enero). *Desarrollo de la habilidad visomotora: Un estudio con el Bender-II* [Development of visual-motor skills: A study with Bender-II]. Trabajo presentado en el Encuentro Científico Internacional 2013 de verano. Lima, Perú.

Rae, G. & Hyland, P. (2001). Generalizability and classical test theory analysis of Koppit'z Scoring System for human figure drawings. *British Journal of Educational Psychology, 71,* 369-182.

Rhodes, R.L., D'Amato, R.C. & Rothlisberg, B.A. (2009). Utilizing a neuropsychological para-

digm for understanding educational and psychological tests. En C.R. Reynolds & E. Fletcher-Janzen (Eds.), *Handbook of Clinical Child Neuropsychology* (3ra. ed., pp. 321-348). New York: Plenum.

Shrout, P.E. & Fleiss, J.L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86,* 420-428. http://dx.doi.org/10.1037/0033-2909.86.2.420

Steiger, J.H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin, 87,* 245-251. http://dx.doi.org/10.1037/0033-2909.87.2.245

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*(2), 615-623. http://dx.doi.org/10.2466/pms.199 0. 70.2.615

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. http://dx.doi.org/10.1177/073428 2909348216

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. http://dx.doi.org/10. 1037/0735-7028.27.3.316

Webber, L.S. & McGillivray, J.A. (1998). An Australian validation of the Kaufman Brief Intelligence Test (K-BIT) with adolescents with an intellectual disability. *Australian Psychologist, 33*(3), 234-237.

Williams, E.J. (1959). The comparison of regression variables. *Journal of the Royal Statistical, 21,* 396-399. http://dx.doi.org/10.1080/00050 069808257412

Zayerz, C. & Manco, C. (2008, Julio). *La ilusión de la igualdad: El caso de la adaptación de pruebas de inteligencia* [Equality illusion: The case of the adaptation of intelligence tests]. Trabajo presentado en el Encuentro Científico Internacional ECI-2008 (Invierno). Lima, Perú.

Instituto de Investigación de Psicología
Universidad de San Martín de Porres
Lima - Perú
Department of Education & Allied Studies
John Carroll University
United States

Fecha de recepción: 24 de septiembre de 2012
Fecha de aceptación: 5 de septeimbre de 2013