Apuntes de Psicología
2014, Vol. 32, número 1, págs. 49-56.
ISSN 0213-3334

Colegio Oficial de Psicología de Andalucía Occidental,
Universidad de Cádiz, Universidad de Córdoba,
Universidad de Huelva, Universidad de Sevilla

Análisis de ítems del nuevo Test Gestáltico Visomotor de Bender (2ª versión)

César MERINO SOTO
Universidad de San Martín de Porres. Lima (Perú)
Ryan A. ALLEN
John Carroll University. Cleveland (Ohio, USA)

Resumen

La segunda versión del test Gestáltico Visomotor de Bender, Bender-II, tiene cambios estructurales y funcionales que lo presentan como un mejoramiento de las versiones anteriores. Sin embargo, hasta la fecha no hay estudios publicados en habla hispana que repliquen sus propiedades psicométricas en el nivel de los ítems. La presente investigación examinó las características psicométricas de los ítems en niños preescolares entre 4 y 5 años de edad. Los resultados indican una elevada sensibilidad de los trece diseños a las diferencias de edad, así como una progresión de la dificultad correspondiente con el orden de presentación. El acuerdo inter-calificadores fue aceptable y los ítems presentaron un apropiado ajuste al TRI modelo *rating scale*; la dimensionalidad fue satisfactoria, sin embargo, hubo errores correlacionados en algunos ítems. Se discuten las características estadísticas de los ítems y algunas diferencias halladas en relación a la muestra de estandarización americana.

Palabras clave: Test de Bender, psicometría, validez, evaluación, niños.

Abstract

The second version of the visual-motor Gestalt test Bender (Bender-II), has structural and functional changes, representing an improvement from previous versions. However, to date there are no published studies that replicate their psychometric properties of the items. The present research examined the psychometric characteristics of the items, in a preschool children between 4 and 5 years old. The results indicate a high sensitivity of designs to age differences, as well as a progression of difficulty corresponding to the order of item presentation. The inter-rater agreement was acceptable, and there is an appropriate adjustment to IRT rating scale model; the dimensionality was satisfactory, however, there is some correlated errors between some items. Statistical characteristics of the items and some differences found with respect to the sample of U.S. standardization are discussed.

Key words: Bender Test; Psychometry; Validity; Assessment; Children.

Desde que Lauretta Bender creó el Test Gestáltico Bender (TGB; Bender, 1938, 1946), éste continuó como una de las medidas más populares utilizadas por profesionales e investigadores (Archer & Newsom, 2000; Piotrowski, 1995; Sullivan & Bowden, 1997). Posteriormente, la creación de muchos sistemas de calificación para el TGB motivó un interés exponencial por estudios de validez y confiabilidad, incrementándose éstos por los años 90 del pasado siglo (Cummings, Hoida, Machek & Nelson, 2003). Actualmente el más importante cambio en el TGB ha sido la segunda edición (Bender-II, Brannigan & Decker, 2003). Esta nueva edición ha incorporado sustanciales cambios estructurales y funcionales que lo pueden señalar como una verdadera nueva versión. Uno de sus cambios es un nuevo método de calificación, del que se deriva un puntaje ordinal luego de evaluar la exactitud de la reproducción. Este método se diferencia de los sistemas discretos de calificación basados en errores, cuestionados por a) su dificultad para interpretar el cambio evolutivo y la eficacia de intervenciones (Decker, 2007), b) por su inestabilidad en su efectivo papel diagnóstico para predecir el rendimiento en lectura (Lesiak, 1984), y c) por no capturar apropiadamente el constructo de habilidad visomotora (Bender, 1965, 1970; Brannigan & Brunner, 2002).

Hasta la fecha, la información sobre el funcionamiento estadístico de los ítems del Bender-II es inexistente en una

Dirección del primer autor: Facultad de Ciencias de la Comunicación, Turismo y Psicología. Instituto de Investigación de Psicología, Universidad de San Martín de Porres. República de Panamá, Tomas Marsano, Lima 15036 (Perú). *Correo electrónico:* sikayax@yahoo.com.ar

Recibido: mayo de 2013. *Aceptado:* febrero de 2014.

muestra independiente de lo reportado en el manual. Excepto el estudio independiente en niños autistas (Volker *et al.*, 2010) y en una muestra hispana (Merino, 2012), los estudios adicionales con el Bender-II han hecho un análisis secundario de la muestra de estandarización (Allen & Decker, 2008; Decker, 2007; Decker, Allen & Choca, 2006); pero en ambas situaciones, se usaron los puntajes como unidades de análisis. Parece que el uso del Bender-II para la investigación en equipos de investigación no anglosajones no ha llamado la atención seriamente como para iniciar estudios psicométricos que investiguen todas sus propiedades psicométricas. Una aparente resistencia al cambio (Brannigan & Decker, 2006) o la creencia que las propiedades psicométricas de las antiguas versiones de pueden transferir al nuevo Bender-II podrían explicar su infrecuente uso.

La información sobre los ítems es una fuente importante para comprender el funcionamiento del puntaje total, y se recomienda que sea bien documentado en relación al constructo y su equivalencia (*American Educational Research Association, American Psychological Association & National Council for Measurement in Education*, 1999; Clark & Watson, 2003). Por lo tanto, la presente investigación aborda el análisis psicométrico de los ítems (diseños) del Bender-II, en una muestra de niños preescolares urbanos. La evaluación de niños preescolares favorece la una temprana identificación de capacidades y dificultades para proveer servicios oportunos de intervención (Beery & Beery, 2000); y la habilidad visomotora es uno de los constructos de mayor impacto y predictividad en las futuras habilidades de aprendizaje (Brannigan & Brunner, 2002; Kulp, 1999; Beery, 2000). Inclusive, aún hasta aproximadamente los 10 años de edad, la habilidad visomotora tiene impacto en la variabilidad de medidas estandarizadas de rendimiento académico (Kulp *et al.*, 2004; Sortor & Kulp, 2003).

El objetivo del estudio de examinar las propiedades psicométricas de los ítems del Bender-II, en relación a parámetros de dificultad, discriminación, dimensionalidad, confiabilidad y de validez de la estructura interna. Este aporte permite evaluar, por otro lado, la replicabilidad de lo reportado en el manual (Brannigan & Decker, 2003) en algunas de estas propiedades, y extender sus hallazgos con aspectos no evaluados y/o reportados en la muestra de estandarización. Esta nueva información sobre los ítems permite conocer el funcionamiento estadístico de los nuevos ítems del Bender-II, característica resaltante del instrumento.

Método

Participantes

Fueron 173 niños de ambos sexos (83 niñas, 48%) distribuidos similarmente en las edades de 4 (78, 45.1%) y 5 años de edad, todos procedentes de un centro educativo privado de educación regular, ubicado en la zona urbana de Lima Metropolitana. Todos los niños de nuestro estudio asistían regularmente a sus clases diarias, y provinieron de un ambiente familiar de nivel socioeconómico medio, y generalmente viviendo con ambos padres. Los padres principalmente se desempeñan como empleados en trabajos estables, y las madres mayormente se desempeñan como amas de casa y algunos trabajos independientes. La mayoría de las familias residen a no más de 2 kilómetros alrededor de colegio. De acuerdo a la información recogida por las autoridades escolares, 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 colegio.

Instrumento

Test Gestaltico Visomotor de Bender, 2da. Versión (Bender-II, Brannigan & Decker, 2003).

La nueva versión del TGB está diseñada para evaluar el funcionamiento visomotor en sujetos entre 4 y 85 años. Contiene 16 láminas apropiadas para extender el rango de aplicación. El Bender-II se aplica en dos fases (copiado de diseños y recuerdo de los diseños dibujados), más dos pruebas complementarias que evalúan la motricidad fina y la percepción visual. En la parte de copia, al sujeto debe reproducir todos los diseños presentados uno por uno; inmediatamente después, se solicita que recuerde los diseños presentados y los dibuje también uno por uno. Complementariamente, se le solicita resolver ítems de coordinación motora y discriminación visual, que sirven como medidas de despistaje de problemas asociados al desempeño visomotor evaluado en la fase de copiado. Para la calificación de cada diseño reproducido en la fase de copia y en la de recuerdo, se aplica el *Sistema Global de Calificación (SGC)*, creado específicamente para esta versión, en que se asigna un puntaje que varía entre 0 y 4 dependiendo del grado de exactitud del diseño dibujado comparado con los ejemplos del manual. El TGB-II respalda sus aspectos normativos y psicométricos desde una muestra de estandarización de más de 4000 personas estratificadas por etnia, educación y estatus socioeconómico, empatando estas características con el censo de la población americano del año 2000 (Brannigan & Decker, 2003). Un reciente estudio en Perú sobre la confiabilidad mostró resultados satisfactorios respecto a la consistencia interna y el acuerdo intercalificadores (Merino, 2012).

Procedimiento

De acuerdo a los planteamientos aceptados para el proceso de adaptación de pruebas psicológicas (Muñiz,

Elosua & Hambleton, 2013), se efectuaron procedimientos respecto a la traducción, aplicación y calificación. Como todo el material del Bender-II está en inglés, se procedió a traducir sus instrucciones de aplicación y calificación. 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 amplia experiencia en el uso del Bender-II, verificó y aprobó la traducción. Todos acordaron que las instrucciones traducidas de aplicación y calificación capturaban el original en inglés.

La administración del Bender-II se hizo en condiciones estandarizadas y manteniendo en lo posible las instrucciones de administración grupal o individual para maximizar la varianza relacionada con el constructo medido (Lee, Reynolds & Willson, 2003; Bracken, 2007), que incluía las condiciones de administración y la secuencia de aplicación de las subpruebas del Bender-II. Antes de aplicar los métodos estadísticos para los objetivos planteados, se verificaron la precisión de la información recolectada. Para obtener una mejor precisión en los puntajes, los protocolos fueron calificados por los tres examinadores que participaron en la recolección de datos, y el promedio de estas calificaciones (con redondeo hacia arriba para permitir solo números enteros) se usaron en los análisis. En segundo lugar, se examinaron las distribuciones para detectar posibles valores extremos, sin embargo ningún dato fue considerado anómalo.

Análisis

El análisis de ítems consistió en explorar varios parámetros; uno de ellos fue la dispersión y la respuesta promedio. Ésta última es la investigación del parámetro de dificultad de los ítems desde el marco de la Teoría Clásica de los Tests. Usando este parámetro, primero se probó la hipótesis de la influencia de la edad, aplicando la prueba *t de Student* para muestras independientes y estimaciones de magnitud del efecto (Coe & Merino, 2003) para ver las posibles diferencias en cada ítem respecto a los dos grupos de edad evaluados (4 y 5 años). En segundo lugar, se hizo el contraste de la hipótesis sobre el ordenamiento de los ítems, es decir, de un decremento lineal a su nivel de dificultad. Para ello se aplicaron contrastes ortogonales a priori, estableciendo como modelo de variación la tendencia lineal. Por otro lado, para obtener los parámetros del dificultad desde el modelo TRI (teoría de respuesta al ítem), los datos se ajustaron a un modelo politómico mediante el programa *Conquest* (Wu, Adams, Wilson & Haldane, 2007); este modelo es más apropiado para el SCG, primero porque a diferencia de otros sistemas de puntuación discretos (por ejemplo, Koppitz, 1984; Sugar, 1995; Watkings, 1976), en el SCG cada ítem se califica entre 0 y 4. En segundo lugar, es un modelo parsimonioso que asume similar poder discriminatorio de los ítems, característica previamente reportada en Merino (2009) y en los estudios normativos originales (Brannigan & Decker, 2003; Decker, 2007). Por lo tanto, el ajuste TRI se hizo con el modelo *rating scale* (Andrich, 1978, 1988).

La exploración de la discriminación se efectuó calculando la correlación ítem-test corregido (Nunnally & Bernstein, 1995). La evaluación estructural de los ítems con el constructo de habilidad de integración visomotora se hizo ajustando los ítems a una estructura latente unidimensional, que es consistente teóricamente con la interpretación de un puntaje único para el Bender-II (Brannigan & Decker, 2003); para ello se aplicó un análisis factorial confirmatorio con AMOS (Arbuckle, 2009), usando el estimador máxima verosimilitud y correlaciones Pearson entre los ítems. Finalmente, para evaluar el acuerdo intercalificadores, se seleccionaron 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 de desarrollo psicomotor. Para este objetivo, se evaluó específicamente el acuerdo inter-calificadores mediante la aplicación de un modelo del análisis de varianza basado en el coeficiente de correlación intraclase, ICC (Shrout & Fleiss, 1979. Este modelo asume que, primero, los calificadores son seleccionados aleatoriamente de una población de potenciales calificadores; segundo, cada calificador evalúa a cada diseño; y tercero, que los objetos de medición (en este caso, los diseños del Bender-II) provienen de una población potencial de estímulos. El modelo de efectos aleatorios de dos vías es apropiado para la mayoría las situaciones de acuerdo inter-calificadores (McGraw & Wong, 1996).

Resultados

Variabilidad

La dispersión de los puntajes en los diseños observó un patrón reconocible de entre las edades (tabla 1). En los cinco primeros ítems, la dispersión fue menor entre los niños de 5 años; esto se relaciona con la facilidad de los primeros ítems para los niños de 5 años, quienes respondieron con más frecuencia puntajes entre 3 y 4, es decir, reproducciones más exactas. En el segundo grupo de ítems, en que la variabilidad fue mayor para los niños de 5 años, los ítems fueron más difíciles para los niños de 4 años y sus respuestas se concentraron en el extremo inferior del puntaje (entre 0 y 1).

Dificultad

En la tabla 1 se muestran las medias para cada diseño. En primer lugar, se examinó la hipótesis de diferencia entre-grupos, teniendo como factor fijo a la edad (4 y 5 años). El grupo de 5 años se desempeñó consistentemente mejor en todos los diseños, y las diferencias estuvieron alrededor de

Tabla 1. Promedio, variabilidad y diferencias estandarizadas para los diseños del Bender-II.

Diseño	4 años (n: 78)		5 años (n: 95)		*d* Cohen	Total	
	M	*DE*	*M*	*DE*		*M*	*DE*
1	2.79	1.05	3.24	0.80	0.49	3.01	0.99
2	1.98	0.84	2.48	0.45	0.78	2.18	0.75
3	1.58	0.90	2.17	0.69	0.75	1.82	0.87
4	1.92	0.95	2.40	0.48	0.68	2.12	0.82
5	0.73	0.75	1.35	0.74	0.83	0.99	0.81
6	0.63	0.65	1.00	0.83	0.49	0.78	0.75
7	0.42	0.67	0.81	0.82	0.51	0.58	0.76
8	0.14	0.28	0.52	0.65	0.72	0.30	0.50
9	0.70	0.73	1.41	0.67	1.02	0.99	0.79
10	0.28	0.61	0.90	0.85	0.81	0.53	0.78
11	0.69	0.63	1.20	0.66	0.79	0.90	0.69
12	0.32	0.47	0.72	0.73	0.63	0.49	0.62
13	0.75	0.57	1.22	0.68	0.74	0.94	0.66

M: media. DE: desviación estándar. *d Cohen*: magnitud del efecto.

lo que puede considerarse una moderada o larga magnitud entre ellas, de acuerdo a las sugerencias sobre el uso de las diferencias estandarizadas (Cohen, 1992). La diferencia entre el grupo de 4 y 5 años en todos los diseños fue estadísticamente significativa (*t de Student* > 3.0), aún con la corrección Bonferroni al nivel *p* establecido (0.05). Respecto a la hipótesis de si la dificultad de cada diseño tenía una progresión de tendencia lineal (es decir que el ordenamiento de los ítems sigue un patrón del más fácil al más difícil), se halló una tendencia lineal de fuerte magnitud, $F_{(1,171)}$=1334, p<0.001, h^2= 0.88. Se detectaron también otras tendencias no lineales también, como la tendencia cuadrática ($F_{(1,171)}$= 624.96, p<0.01, h^2 =0.78) o polinómica de orden 11 ($F_{(1,171)}$ = 201,50, p<0.01, h^2 =0.54), ambos de menor magnitud.

En la tabla 2 aparecen los estadísticos de ajuste para los ítems con el modelo *rating scale*. Al transformar los estadísticos a variables *t* (Wu, Adams, Wilson & Haldane, 2007), ninguno de ellos fue estadísticamente significativo (< 0.196), excepto los diseños 1, 5 y 13 en ambas medidas de ajuste (INFIT y OUTFIT). Estos indicadores de ajuste generalmente estuvieron dentro del rango recomendado (0.6-1.4) para evaluar el ajuste en ítems politómicos al modelo de comparación (Wang & Cheng, 2005; Wright & Linacre, 1994). Los valores INFIT y OUTFIT alrededor de 1.0, indicando que apenas existe entre el 1% y 2% de ambigüedad en el modelo aplicado en cada ítem (Wright & Linacre, 1994). Estos resultados indican que el ajuste al modelo TRI elegido es una aceptable. Para comparar los parámetros θ con los reportados en la muestra de estandarización (Decker, 2007), se crearon intervalos de confianza al 95% con el error estándar de cada q reportado en la tabla 2. Las diferencias estadísticamente significativas se hallaron en los diseños del 2 al 4, y del 9 al 13; para el resto de los diseños, los parámetros fueron de similar magnitud a lo reportado en Decker (2007). En el primer grupo de estos diseños (2 al 4), el parámetro q fue de mayor magnitud; en el segundo grupos de ítems, la muestra de estandarización americana obtuvo mayores parámetros.

Tabla 2. Parámetros psicométricos para los 13 diseños del Bender-II.

	RITC			ICC			Modelo rating scale			
	4a	*5a*	*Total*	*4a*	*5a*	*Total*	θ *(error estándar)*	*INFIT*	*OUTFIT*	*CFA*
Diseño 1	0.40	0.16	0.37	0.82	0.71	0.79	-3.81 (0.08)	**1.92***	**2.00***	0.40
Diseño 2	0.52	0.15	0.50	0.76	0.29	0.68	-1.91 (0.07)[a]	0.87	0.89	0.54
Diseño 3	0.49	0.20	0.49	0.77	0.57	0.73	-1.22 (0.07)[a]	1.07	1.10	0.53
Diseño 4	0.56	0.31	0.54	0.84	0.53	0.79	-1.76 (0.07)[a]	0.86	0.91	0.59
Diseño 5	0.65	0.48	0.66	0.75	0.68	0.75	0.38 (0.07)	**0.77***	**0.77***	0.74
Diseño 6	0.38	0.51	0.48	0.56	0.63	0.61	0.81 (.07)	**1.29***	1.12	0.46
Diseño 7	0.42	0.57	0.52	0.78	0.75	0.77	1.22 (0.08)	1.09	1.12	0.54
Diseño 8	0.36	0.55	0.52	0.35	0.52	0.52	2.26 (0.08)	**0.76***	0.89	0.52
Diseño 9	0.46	0.38	0.56	0.70	0.60	0.70	0.32 (0.07)[b]	0.93	0.95	0.62
Diseño 10	0.37	0.52	0.54	0.82	0.71	0.78	1.29 (0.08)[b]	0.97	1.16	0.56
Diseño 11	0.46	0.32	0.51	0.68	0.52	0.65	0.50 (0.07)[b]	0.91	0.88	0.57
Diseño 12	0.44	0.45	0.51	0.52	0.68	0.64	1.44 (0.08)[b]	0.86	0.93	0.57
Diseño 13	0.53	0.47	0.58	0.70	0.70	0.71	0.45 (0.27)[b]	**0.67***	**0.63***	0.66

RITC: correlación ítem-test corregida. ICC: correlación intraclase. θ : estimación de habilidad latente. CFA: cargas obtenidas del análisis factorial confirmatorio (CFA: confirmatory factor analysis), función máxima verosimilitud.
a: θ en la muestra peruana es mayor. b: θ en la muestra de estandarización (Decker, 2007) es mayor.

Discriminación

Las correlaciones ítem-test (r_{itc}) para la muestra total superaron el valor 0.30; pero las magnitudes no fueron constantes entre los grupos de edad (tabla 2). En el grupo de 5 años, los primeros tres ítems tuvieron niveles debajo del criterio, y la mayoría de los ítems tuvieron mejor discriminación en los participantes de 4 años. La baja discriminabilidad de los tres primeros diseños parece asociarse a su menor grado de dispersión en la muestra de 5 años. Al comparar la magnitud de las r_{itc} entre las edades, la máxima diferencia ocurrió en el diseño 2, que no fue estadísticamente significativas ($Z = 2.72$, $p = 0.003$) luego de aplicar la corrección Bonferroni, (0.05/13 = 0.0038).

Dimensionalidad

Se hizo un análisis previo del número de factores subyacente a los ítems, mediante el gráfico *scree test* (Cattell, 1966) y el *análisis paralelo* (Horn, 1965). Ambos procedimientos unívocamente sugirieron que una sola dimensión en suficiente para explicar la varianza de los ítems del Bender-II. Los resultados de aplicar el método CFA indican que las cargas factoriales fueron elevadas ($\lambda \geq 0.40$) e indican un elevada varianza entre los diseños y el atributo latente (tabla 2). Sin embargo, el modelo inicial puede considerarse marginalmente satisfactorio: χ^2 (gl=65) = 152.2, $p < 0.01$, RMSEA = 0.08 (IC 90% = 0.07, 0.10), CFI = 0.86, χ^2/gl = 2.34.

Se procedió a revisar el modelo mediante dos pasos: un enfoque estadístico y otra lógica-teórica. En el primer criterio, la observación de los índices de modificación reveló que los cambios estadísticos más significativos corresponden a la covariación entre los errores de los ítems 6 y 7, y 6 y 8. En segundo lugar, estos ítems son cualitativamente similares, pues usan patrones de puntos como estímulos, y algunos recientes estudios han mostrado que la covariación entre estos ítems es mayor comparado a los demás ítems del Bender (Liz & Mazzeschi, 2000; Merino, 2009). Por lo tanto, el ajuste final se obtuvo imponiendo covariación entre los errores de los diseños 6-7, y 6-8. La consecuente magnitud del ajuste fue excelente, tal como es recomendado para modelos unidimensionales (Reeve *et al.*, 2007): χ^2 (gl=63) = 103.26, $p = 0.001$, RMSEA = 0.01 (IC 90% = 0.03, .08), CFI = 0.93, χ^2/gl = 1.63.

Debe anotarse que la magnitud de las cargas factoriales fueron menores a las reportadas en la muestra de estandarización americana en Brannigan y Brunner (2003).

Discusión

El objetivo de la investigación aquí presentada fue evaluar las cualidades psicométricas de los ítems de la nueva versión del Test Gestáltico Visomotor de Bender, Bender-II. Originalmente, esta versión tiene un fuerte respaldo psicométrico en la muestra de estandarización americana, que incluyó una muestra de hispanos inmigrantes proporcionalmente distribuidos de acuerdo a los datos censales americanos. Considerando la variabilidad ha sido consistente la dificultad de los diseños en ambas edades. Los ítems que requerían más habilidad o menos habilidad, mostraron una menor variabilidad. Aunque se ha observado que en dos diseños no ocurre bondad de ajuste con el modelo *rating scale*, estos han sido razonables considerando el tamaño muestral en los extremos de la distribución, y la restricción del rango de respuesta en estos diseños.

El análisis de la discriminación de los ítems y la relación estructural con el atributo latente pueden considerarse como satisfactorios, aunque se hallaron diferencias con las estimaciones obtenidas del estudio original. Por ejemplo, las cargas factoriales halladas fueron más bajas que las reportadas en la muestra de estandarización de mil niños entre 4 y 7 años (Brannigan & Brunner, 2003), en donde que las cargas fueron mayores de 0.70. En nuestros resultados las cargas son mayores de 0.40, lo que sugiere una mayor introducción de varianza única proveniente la interacción de factores aleatorios o de habilidades específicas que explicaron un parte del desempeño de los niños muestreados.

Las cargas obtenidas en nuestro estudio se derivaron mediante el ajuste empleado para mejorar la representación del modelo unidimensional, estrategia impuesta debido a la covariación entre ítems figurativamente similares (diseños de puntos), y puede explicar esta varianza única. Esta la modificación estructural efectuada sobre algunos de los ítems sugiere un rendimiento diferencial y único en una parte del Bender-II. Estos resultados dejan abierta una línea de futura investigación para evaluar si la mayor covariación entre los diseños 6, 7 y 8 es idiosincrásica a la muestra o es un fenómeno replicable. Otra explicación es la presencia de una dimensión adicional que aportaría una mejor explicación a la interpretación de los puntajes en el Bender-II; sin embargo, es más plausible pensar en un rendimiento único en la muestra de estudio, y que fue particularmente sensible a responder a este grupo de ítems posiblemente involucrando similares conductas para reproducir tales diseños.

Aunque el rendimiento en estos ítems podría modelarse con un factor adicional, no se hizo este ajuste pues no es teóricamente compatible con la unidimensionalidad latente del TGB. Es plausible que la modificación estructural efectuada sobre los ítems que requieren la reproducción de patrones de puntos sugiera un rendimiento diferenciado del resto de los ítems. Esto parece ser congruente con algunos hallazgos en que se aplicaron análisis factoriales con otros sistemas de calificación (Guertin, 1954; Roberds, 1974; Sisto, Noronha & Santos, 2005), donde se tendía a extraerse más de un factor. Debido que el nuevo sistema de calificación (Sistema de Calificación Global) se basa en la exactitud reproductiva de

los diseños, el puntaje general operacionalizaría la contribución de todos los ítems en esta interpretación, como lo se planteaba años anteriores para otro sistema de calificación (Wagner & Marsico, 1991)

Por ejemplo, la discriminación de los ítems no pueden ser comparada con el estudio original (Brannigan & Brunner, 2003), que no lo reporta; pero en términos absolutos, la capacidad discriminativa se los ítems ha sido satisfactoria considerando la muestra total, aunque algo variable entre los dos grupos de edad examinados. En conjunto, sin embargo, la capacidad discriminativa de los ítems favorecería el uso del puntaje total como un indicador para la detección de desempeños en un amplio rango de distribución de puntajes. Esta situación se empareja con la heterogénea dificultad de los diseños, que favorece también la discriminación en diferentes puntos de la distribución de puntajes (Anastasi & Urbina, 1998).

Por otro lado, los parámetros hallados desde el modelo TRI (Andrich, 1978; 1988) sugieren que se puede lograr una medición más precisa (Adams & Khoo, 1993; Smith *et al.*, 1998), especialmente con la inclusión de cuatros ítems nuevos para extender el escalamiento del Bender-II hacia desempeños bajos en referencia al rango total de puntajes en el Bender-II. La aplicación del modelo *rating scale* parece ser consistentemente efectiva cuando de aplica al Bender-II, pues este modelo también tuvo un ajuste apropiado en la muestra de estandarización original. Otro patrón que se pudo observar fue que en los últimos 6 diseños, los parámetros θ fueron menos variables que en el resto de los diseños, y las distancias entre los θ fueron algo más amplias. Con estos nuevos ítems, y una mayor dispersión de las dificultades, el mejoramiento de la evaluación por medio del TGB se haría más efectivo en los niños pequeños y en los niños con déficits cognitivos, haciendo que el suelo del test de extienda para poder discriminar aquellos con rendimiento muy bajos (Bracken, 1987). Estos resultados dan un mejor respaldo para el uso de esta nueva versión en la muestra de estandarización y que puede extenderse similarmente en muestras independientes.

En la comparación con la muestra de estandarización americana, las estimaciones θ para los ítems mostraron diferencias entre los primeros y últimos ítems. En los niños peruanos el primer grupo de ítems mostró ser más fácil para los niños americanos, mientras que para el grupo de los últimos ítems, estos mostraron ser más fáciles para los niños americanos. Este hallazgo debe tomar en cuenta que la composición de la muestra americana en esta comparación incluyó un mayor rango de edad, mientras en la muestra peruana solo fueron niños entre 4 y 5 años, y en un tamaño muestral mucho menor. Estas diferencias pueden sugerir crear normas diferentes en la interpretación del puntaje total. La información acumulativa que proporcionan los ítems al puntaje total permite valorar psicométricamente la interpretación de los resultados.

Los resultados obtenidos deben, sin embargo, ser evaluados en el contexto de las limitaciones del estudio. Dado que el tamaño muestral compromete el poder estadístico en el presente estudio y la estabilidad de los resultados hallados, se requiere resolver esto antes de asegurar las conclusiones sobre las características a nivel del ítem del Bender-II. Por otro lado, debido al muestreo incidental, se requiere un diseño muestral probabilístico será necesario para poder obtener parámetros más representativos y similares a la población de referencia. Un aspecto no evaluado en el presente estudio fue el sesgo o funcionamiento diferencial de los ítems, lo que puede evaluarse entre varones y mujeres; sin embargo, que no hay motivos para pensar que los ítems pueden funcionar sesgadamente hacia uno de los sexos, considerando que el estudio de su construcción tomó en cuenta esto (Brannigan & Decker, 2003). Finalmente, el estudio de la confiabilidad intercalificadores puede adquirir mayor poder descriptivo si se aplica la teoría de la generalizabilidad, que sería abordado en una posterior investigación.

Junto a otro reciente estudio en muestra peruana (Merino, 2012), la investigación presentada aquí contribuye a acumular evidencia sobre esta nueva versión, la que aún es poco incluida en la investigación del desarrollo infantil en el mundo hispano, y por lo tanto no se puede hallar numerosa bibliografía actualizada. Hace más de 40 años, Bender (1965) relacionaba el apropiado uso de su prueba con la totalidad de la conducta del niño durante la ejecución de su prueba; la presente segunda versión es coherente con esa idea. Sin embargo, sin el respaldo psicométrico en grupos interculturalmente diferentes no se podría confiar en que sus resultados sean apropiados (*American Educational Researsh Association et al.*, 1999).

Referencias

Adams, R.J., & Khoo, S.T. (1993). *Quest: The interactive test analysis system* [Computer program manual]. Hawthorn: The Australian Council for Educational Research.

Allen, R.A., & Decker, S.L. (2008). Utility of the Bender Visual-Motor Gestalt Test - Second Edition- in the assessment of attention-deficit/hyperactivity disorder. *Perceptual and Motor Skills*, *107*, 663-675.

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

Anastasi, A., & Urbina, S. (1998). *Tests psicológicos*. México: Prentice Hall.

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

Andrich, D. (1988). *Rasch models for measurement*. Newbury Park: Sage.

Arbuckle, J.L. (2003). *Amos 18* [Computer program]. Chicago: SPSS.

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

Beery, K.E., & Beery, N.A. (2000). *Prueba Beery-Buktenica del desarrollo de la integración visomotriz* (4ª Ed.). México: El Manual Moderno.

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

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

Bender, L. (1965). On the proper use of the Bender Gestalt Test. *Perceptual and Motor Skills, 20*, 189-190.

Bender, L. (1970). Use of the Visual Motor Gestalt Test in the diagnosis of learning disabilities. *Journal of the Special Education, 4, 29-39*.

Bracken, B.A. (1987). Limitations of preschool instruments and standards for minimal levels of technical adequacy. *Journal of Psychoeducational Assessment*, *4*, 313-326.

Bracken, B.A. (2007). Creating the optimal preschool testing situation. En B.A Bracken & R.J. Nagle (Eds.), *Psychoeducational Assessment of Preschool Children* (pp. 137-153). Mahwah: 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: Charles Thomas.

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

Brannigan, G.G., & Decker, S.L. (2003b). Bender-Gestalt II. *American Journal of Orthopsychiatry, 76*(1), 10-12.

Brannigan, G.G., Decker, S.L., & Madsen, D.H. (2004). Innovative features of the Bender-Gestalt II and expanded guidelines for the use of the Global Scoring System. *Bender Visual-Motor Gestalt Test, Second Edition Assessment Service Bulletin, 1*. Itasca: Riverside Publishing.

Cattell, R.B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*, 245-276.

Clark, L.A., & Watson, D. (2003). Constructing validity: Basic issues in objective scale development. En A.E. Kazdin (Ed.), *Methodological issues and strategies in clinical research (3ª Ed.)* (pp. 207-231). Washington: APA

Coe, R., & Merino, C. (2003). Magnitud del efecto: Una guía para investigadores y usuarios. *Revista de Psicología PUCP, 21*(1), 147-177.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155-159.

Cummings, J.A., Hoida, J., Machek, M.G., & 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 (2nd. Ed.)*, pp. 498-518. Nueva 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 Wechsler Intelligence Scale for Children-III. *Perceptual and Motor Skills*, *102*, 133-141.

Guertin, W.H. (1954). A factor analysis of curvilinear distorsions on the Bender-Gestalt. *Journal of Clinical Psychology, 10,* 12-17.

Horn, J.L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30, 179-185.

Kulp, M.T. (1999). Relationship between visual motor integration skill and academic performance in kindergarten through third grade. *Optometry and Vision Science, 76*, 159-163.

Kulp, M.T., Earley, M.J., Mitchell, G.L., Timmerman, L.M., Frasco, C.S., & Geiger, M.E. (2004). Are visual perceptual skills related to mathematics ability in second through sixth grade children? *FOCUS on Learning Problems in Mathematics, 26*(4), 44-51.

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

Lesiak, J. (1984). The Bender Visual Motor Gestalt Test: Implications for the diagnosis and prediction of reading achievement. *Journal of School Psychology, 22*(4), 391-405.

Lis A., & Mazzeschi, C. (2000) The Bender Gestalt Test in an Italian sample: an analysis of Koppitz developmental bender scoring system deviation. *Perceptual & Motor Skills, 90*, 373-385.

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

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

Merino, C. (2012). Confiabilidad en el Test Gestáltico de Bender (2ª versión), en una muestra independiente de calificadores. *Revista de Investigación Educativa, 30*(1), 223-234.

Muñiz, J., Elosua, P., & Hambleton, R.K. (2013). Directrices para la traducción y adaptación de los tests: segunda edición. *Psicothema, 25*(2), 151-157.

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

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

Reeve, B.B., Hays, R.D., Bjorner, J.B., Cook, K.F., Crane, P.K., Teresi, J.A., Thissen, D., Revicki, D.A., Weiss, D.J., Hambleton, R.K., Honghu, L., Gershon, R., Reise, S.P., Lai, J., & Cella, D. (2007). Psychometric evaluation and calibration of health-related quality of life items banks: Plans for the Patient-Reported Outcome Measurement Information System (PROMIS). *Medical Care*, *45*(5 suppl 1), S22-S31.

Roberds, J. (1974). The Bender Gestalt and the Raven Matrices Progressive measures for perceptual behavior, motor behavior, and perceptual-motor behavior. Presentado en el *59th Annual meeting for the American Educational Research Association* (Chicago, Abril).

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.

Smith, R.M., Schumaker, R.E., & Bush, M.J. (1998). Using item mean squares to evaluate fit to the Rasch model. *Journal of Outcome Measurement, 2*, 66-78.

Sortor, J.M., & Kulp, M.T. (2003). Are the results of the Beery-Buktenica Developmental Test of Visual-Motor Integration and its subtests related to achievement test scores? *Optometry and Vision Science, 80*(11), 758-763.

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

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.

Wang, W., & Chen, C. (2005). Item parameter recovery, standard error estimates, and fit statistics of the WINSTEPS program for the family of Rasch models. *Educational and Psychological Measurement, 65*, 376-404.

Wright, B., & Linacre, M. (1994). Reasonable mean-square fit values. *Rasch Transactions*, *8(3)*. [En línea: http://www.rasch.org/rmt/rmt83b.htm].

Wu, M.L., Adams, R.J., Wilson, M.R., & Haldane, S.A. (2007). *ACER Conquest version 2.0: Generalized item response modeling software*. Victoria: ACER Press.