Evaluación de la confiabilidad del Sistema Cualitativo de Calificación para la versión modificada del Test Gestáltico de Bender*

Assessment of the Reliability in Two Age Groups Using the Qualitative Scoring System for the Modified Bender Gestalt Test

Recibido: marzo 13 de 2009 | Revisado: septiembre 3 de 2009 | Aceptado: marzo 29 de 2010

César Merino Soto**
Luis Benites Morales
Universidad de San Martín de Porres, Lima, Perú

Resumen

El presente estudio tiene por objetivo obtener evidencias de confiabilidad por consistencia interna e intercalificadores para el Sistema Cualitativo de Calificación (Brannigan & Brunner, 2002) aplicado al Test Gestáltico de Bender – Modificado. Los participantes fueron 86 niños, divididos en dos grupos de edad preescolar y escolar. Tres estudiantes de pregrado calificaron los diseños de ambos grupos de niños. Los resultados señalan buenos niveles de confiabilidad intercalificadores y de consistencia interna en el grupo de preescolares, mientras que estos niveles fueron más bajos en el grupo de escolares. Estas diferencias establecen la relación entre estos dos aspectos del error de medición, en los puntajes de esta nueva versión de Test de Bender, y el énfasis en un adecuado entrenamiento de medidas que requieren el juicio del examinador. Se discuten los resultados considerando la potencial utilidad de esta relativa nueva versión del Test Gestáltico de Bender para la práctica clínica y de investigación.

Palabras clave autores
Confiabilidad, evaluación, visomotor, test Gestáltico de Bender, psicometría.
Palabras clave descriptores
Psicometría, Prueba Gestalt, reproducibilidad de resultados.

Abstract

This study is looking for evidences of reliability, for the Qualification Qualitative System (Brannigan y Brunner, 2002) applied to the Bender Gestalt Test – Modified. The participants were 86 children, divided in two groups: pre- school and school; and three students who scored the designs in both groups. The analysis was done in the final grade and the item. The results pointed to the good levels of results of external reliability and internal consistence in the pre- school group, while these levels were scored in the school group. These differences establish the relation between these two aspects of measurement error and the emphasis in an appropriate training of measurements that require the examiner's judgments. We discussed our results considering the potential utility of this relative version of the Bender Gestalt Test for the clinical practice and investigation as well.

Key words authors
Reliability, assessment, visualmotor, Bender Gestaltic Test, psychometry.
Key words plus
Psychometrics, Bender Gestalt Test, reproducibility of results.

Para citar este artículo. 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), 237-249.

* Artículo de investigación.

** Escuela Profesional de Psicología; Av. Tomás Marsano 242, Lima 34, Perú. E-mails: cmerino@usmp.edu.pe; lbenites1@usmp.edu.pe

El Test Gestáltico de Bender (TGB) (Bender, 1987) continúa siendo, desde hace varias décadas, una de las pruebas más populares y frecuentemente administradas, y su uso ha generado más de 1 000 artículos de investigación (Brannigan & Decker, 2003). Hasta la fecha, se observan muchos sistemas de calificación, para niños y adultos, incluyendo las más actuales y psicométricamente robustas, como el nuevo Sistema de Calificación Global de Brannigan y Decker (2003) y el Sistema de Calificación Evolutiva de Koppitz - II (Reynolds, 2007). Las diferentes versiones de los sistemas de calificación se han sometido a evaluación de su validez de constructo y del impacto del error de medición, particularmente de la variabilidad del calificador. En los estudios publicados sobre la confiabilidad de este instrumento, la información sobre el error de medición proveniente de los calificadores se expresa bajo la forma de coeficientes de correlación Pearson. Por ejemplo, usando uno de los sistemas más populares para el TGB aplicados en niños, el acuerdo entre calificadores usando el Sistema Evolutivo de Koppitz (1984) llegaron a coeficientes entre 0.92 y 0.95 adultos con retardo mental (Hustak, Dinning & Andert, 1976). Posteriormente, Aylward y Smidth (1986) hallaron que este sistema es comparable a otras medidas de funcionamiento visomotor con respecto al acuerdo intercalificador, pero que la confiabilidad es variable según el nivel de desempeño del examinado. En niños, más recientemente Rae y Hyland (2001) reportaron coeficientes Pearson mayores de 0.80 entre los calificadores, usando este sistema. Otro sistema de calificación, el Sistema Watkins reportó confiabilidades intercalificadores desde 0.80 hasta 0.97 para el puntaje total (Watkins, 1976). Y estudios posteriores con la escala psicopatológica del Sistema Hutt también reportaron excelentes niveles de acuerdo intercalificadores sobre los diseños de niños escolares entre 7 y 10 años (Rossini, 1993). Otras modificaciones del Bender Gestalt para el despistaje de la disfunción cerebral han reportado también niveles elevados (DeCato & Meldrum, 1989). Elevadas confiabilidades interexaminadores también se han hallado para el nuevo Bender-II (Brannigan & Decker, 2006), que van desde 0.83 hasta 0.94 (mediana = 0.90) y también para el nuevo el Koppitz-II (Reynolds, 2007). Esta breve revisión de los hallazgos sobre el acuerdo entre calificadores sugiere que las diferentes versiones de puntación del TGB son moderadamente variables, pero en un estrecho rango, y estas variaciones tienden a ubicarse satisfactoriamente dentro del espacio aun considerado como apropiado para uso clínico.

La estrategia más común es evaluar la confiabilidad intercalificador con un mínimo de dos calificadores, aunque otras estrategias también se han aplicado en la fase de análisis y en el proceso de calificación. Por ejemplo, Swensson y Hill (1990) obtuvieron los puntajes de 12 calificadores de variada experiencia profesional clínica, sobre cuatro protocolos, hallando correlaciones estadísticamente significativas en los indicadores evolutivos y los emocionales. Un enfoque similar usaron Morsbach, Priori y Firnell (1975) con menos calificadores, pero añadiendo la exploración de la confiabilidad test-retest, medio año después de la primera aplicación. Por lo tanto, el número de calificadores y la evaluación entre e intracalificadores han sido estrategias usadas en la determinación del grado de acuerdo.

Por otro lado, entre los sistemas más usados, el Sistema de Koppitz (1963, 1984) ha permanecido muy popular a través de los años y ha generado más de 300 artículos publicados (Reynolds, 2007) desde su primera publicación en los años 60, pero su eficacia ha sido variable respecto a su validez predictiva. Algunos estudios han reportado que las correlaciones del desempeño en el TGB, usando este sistema con criterios relevantes, han sido de menor magnitud que otros sistemas de calificación. Johnston y Lanak (1985) hallaron un mejor desempeño de las reglas de identificación de déficits visomotores en el Sistema Watkins, en niños referidos a evaluación neuropsicológica. De manera similar, el Sistema Koppitz tiende a explicar menos varianza respecto a criterios de rendimiento escolar estandarizado que otros sistemas más recientes (Brannigan & Brunner, 1989, 1996, 2004; Chang, 2002; Parsons & Weinberg, 1993). Apuntar el in-

terés hacia nuevos enfoques y con mejor respaldo científico es necesario para una apropiada práctica profesional. Pero las pruebas señaladas por diversos autores parecen estar midiendo de la misma manera un atributo, y lo que ocurre es que, aunque estén etiquetados igualmente, las pruebas no deben asumirse como sustitutos intercambiables, porque cada uno puede demandar diferente tiempo para calificar los protocolos, un enfoque diferente de abordaje de la calificación y los niveles de acuerdo intercalificador pueden ser bajos (Preda, 1997). Las consecuencias de todo esto son las divergencias en su validez de constructor, cuando se los usan para evaluar los cambios en programas remediales y para obtener descripciones evolutivas del desarrollo visomotor en general (Palisano & Dichter, 1989).

Hay numerosa investigación sobre los correlatos en el desempeño del Test de Bender, específicamente para el Sistema de Koppitz, pero estos provienen casi totalmente del mundo anglosajón. Por otro lado, una revisión informal de las investigaciones no publicadas a nivel de pregrado y postgrado llevaría a resaltar que la evaluación de las propiedades psicométricas no es uno de sus objetivos principales, si es que acaso se los considera. Por lo tanto, puede ser infrecuente hallar resultados sobre el grado de acuerdo entre calificadores en estas investigaciones no publicadas, a menos que se encuentren bajo una adecuada asesoría y protocolo de investigación. Esto está probablemente asociado a los cursos de pregrado de medición y pruebas que aún permanecen enseñando instrumentos con normas antiguas o sin un eficiente análisis crítico de sus propiedades psicométricas. Específicamente sobre el TGB, aún con la popularidad que tiene este el Sistema Koppitz para niños, hay un consistente vacío para presentar sus propiedades en términos de confiabilidad intercalificador en los estudios que lo usan. Sin embargo, ya que los datos que manejan los investigadores generalmente están a nivel del ítem, hay la posibilidad razonable para hacerlo. Los estándares de información psicométrica propuestos por la American Educational Research Association (AERA), la American Psychological Association (APA) y el National Council on Measurement in Education (NCME) en 1999, recomiendan que se evalúe la confiabilidad de las mediciones en cada contexto de evaluación, y tomar solo como referencia aquellos publicados en los manuales, si las diferencias normativas son verificadas. Junto con esta advertencia, la creciente producción tecnológica en los materiales de evaluación lleva al usuario a utilizar más que un criterio impresionista para elegir una prueba, sino más bien revisar la información psicométrica publicada y considerar su grado de antigüedad.

En el área de la evaluación actual de la habilidad visomotora en niños, uno de los competidores psicométricamente robustos, y aún poco conocido, es el Sistema de Calificación Cualitativa (SCC) de Brannigan y Brunner (1989, 1996, 2002), que se aplica a la versión modificada del TGB. Esta versión fue modificada para niños menores a 8 años, basándose en las sugerencias de la misma Bender sobre la elección de las láminas y el sistema de puntuación (Brannigan & Brunner, 2002) para la batería de Hirsch (Jansky & de Hirsch, 1972). Estudios con la versión modificada del TGB aplicando el SCC para niños preescolares y de primaria, indican que los niveles de acuerdo son elevados (Brannigan & Brunner, 2002; Chang, 2001; Fuller & Vance, 1995). Este evalúa la exactitud de cada reproducción sobre una escala de 6 puntos, y en un rango de 6 a 0. Se diseñó para evaluar la calidad global de las reproducciones de los niños de 4 años y 6 meses hasta 8 años y 5 meses. Este sistema de calificación usa un enfoque estricto de puntuación basado en que el diseño reproducido debe ser tan bueno o mejor que el ejemplo dado en el manual, para recibir la puntuación en el nivel de calidad correspondiente. El sistema se aplica a la versión modificada de la prueba del Bender, que únicamente utiliza 6 diseños que son los más apropiados en la predicción del rendimiento escolar en niños de temprana edad. El eje principal que llama la atención de este enfoque, es el abordaje gestáltico utilizado para evaluar la representación global de cada diseño copiado por el niño (Brannigan & Brunner, 2002), y que, por lo tanto, demanda el juicio del examinador para otorgar el puntaje a cada diseño de acuerdo a su semejanza con el diseño. Esta situación demanda al calificador que

no se concentre en los detalles o partes discretas de los diseños reproducidos, sino más bien sobre su impresión respecto al grado de exactitud con el estímulo presentado. Este aspecto es esencialmente cualitativo y requiere una observación de la gestalt, de tal modo que evita un análisis discreto de los errores en las reproducciones, tal como lo induce el Sistema de Koppitz. Este sistema destaca la evaluación de errores discretos en reproducción de cada una de las 9 láminas, pero ha sido criticado por enfatizar la sobresimplificación y el examen molecular de los errores (Brannigan & Brunner, 2002). Sistemas paralelos se han propuesto (Parsons & Weinberg, 1993; Sugar, 1995) pero no han sido extensamente evaluados psicométricamente y, por lo tanto, han permanecido relativamente en el anonimato, para la comunidad científica.

El presente estudio tendrá como objetivo general evaluar dos fuentes de error de medición: la consistencia interna y el acuerdo intercalificadores (Anastasi & Urbina, 1997). La consistencia interna y el grado de acuerdo se estimarán respecto al uso del Sistema de Calificación Cualitativa, SCC (Brannigan & Brunner, 2002), que se aplica para la versión modificada del Test Gestáltico Visomotor de Bender. El segundo objetivo del trabajo será comparar el acuerdo intercalificadores sobre la calificación de las reproducciones en niños de dos edades diferentes; de nivel preescolar y nivel primario, provenientes de Lima metropolitana. Este último objetivo relaciona el grado de desempeño en el funcionamiento visomotor y el nivel de acuerdo entre calificadores. Este planteamiento emergió del trabajo de Aylward y Smidth (1986), quienes hallaron que los desempeños pobres en la reproducción de las figuras del TGB tendían a producir menor acuerdo entre calificadores; esta afirmación es paralela con la efectuada por Reynolds y Hickman (2004) para la nueva versión de la Prueba del Dibujo de la Figura Humana para la estimación de la inteligencia. Ellos señalaron que los dibujos en niños de menor edad, tienden a producir mayores discrepancias en el uso de criterios que exigen juicio del examinador; su conjetura no presentó datos empíricos ni antecedentes que lo respalden, pero halló que el grado de acuerdo entre examinadores fue similarmente alto ($r > 0.85$) en la calificación de dibujos realizados entre niños y adolescentes.

Método

Participantes

Los participantes son una muestra de 86 niños(as) que forma parte de un estudio de validación y normalización de una prueba de despistaje para habilidades primer grado desarrollado recientemente (Merino, 2006). Los participantes se dividieron en dos grupos de edad provenientes cada uno de instituciones estatales independientes (4 a 5 kilómetros alejados entre sí), pero del mismo sector educativo, en el centro de Lima metropolitana. El nivel 1 estuvo distribuido por 42 preescolares (24 varones, 57.1 %), de 4 y 5 años ($n = 19$, 45.2 %) de edad, procedentes de tres Programas No Escolarizados de Educación Inicial (los PRONOEI); estos programas se orientan a niños de baja condición socioeconómica y son ubicados estratégicamente en cada comunidad. El currículo educativo es el mismo que los centros preescolares escolarizados, pero contienen menos recursos materiales e infraestructurales (Merino, Díaz, Zapata & Benites, 2006). Los niños fueron seleccionados aleatoria y proporcionalmente desde estos tres PRONOEI.

El grupo de niños del nivel 2 estuvo compuesto por 44 escolares, 20 varones (45 %) estudiando en segundo y tercer grado de primaria, en el turno de la mañana de un colegio público ubicado en una zona urbana en un distrito al sur de Lima metropolitana; la edad mínima fue 6 años ($n = 4$, 9.1 %) y la máxima 9 ($n = 1$, 2.3 %); la mayor parte de este grupo tuvo una edad de 6 años ($n = 22$, 50 %) y 7 años ($n = 17$, 38.6 %). Como es usual en los colegios públicos de Perú, cada salón es unidocente y se maneja alrededor de 30 niños por aula. Dada la zona de ubicación del colegio, las familias de los niños en su mayoría alcanzaron el nivel secundario, y las madres tienden a pasar más horas con el niño que el padre, ya que se ocupan del hogar y eventualmente realizan trabajos independientes; y mayoritariamente, las familias de los niños conviven con otros familiares. Por otro lado, los calificadores

fueron tres estudiantes (mujeres) de mitad de la carrera de Psicología, en una universidad estatal en Lima; son de asistencia regular y perteneciente el tercio medio en rendimiento académico. Estos estudiantes no tenían experiencia previa en la administración y calificación de alguna versión del TGB, pero sí en instrumentos de desarrollo psicomotor. Se consideró que la inexperiencia específicamente en algún sistema de calificación para el TGB puede ser una condición facilitadora del aprendizaje de este nuevo sistema, pues la nueva información no tendría que competir con el aprendizaje previo, que invocaría otra manera de abordar la calificación y asignar los puntajes.

Instrumento

Test Gestáltico de Bender Modificado. La versión modificada contiene seis de los diseños originales (A, 1, 2, 4, 6 y 8) para su aplicación a niños preescolares hasta los primeros grados del nivel primario (4.5 hasta 8.5 años), dado que son los más apropiados para niños pequeños. El manual describe un sistema para puntuar el desempeño gráfico del niño, el SCC (Brannigan & Brunner, 2002) de 6 puntos, desde una puntuación de 0 (líneas aleatorias, garabateo, sin concepto del diseño) hasta 5 (representación exacta del diseño) y que logran gran diferenciación en la evaluación de la calidad los dibujos. Esta versión se califica por un método de inspección global, que refleja el grado de diferenciación y de la gestalt de los diseños reproducidos. La investigación sobre la confiabilidad interna, test-retest e intercalificadores (Fuller & Vance, 1995), y la validez del SCC da soporte a sus propiedades métricas y sus cualidades instrumentales en la evaluación psicopedagógica (Brannigan & Brunner, 2002). El SCC acepta que una de las modalidades de administración sea la grupal, ya que se hallan solo diferencias pequeñas entre la administración individual (Caskey & Larson, 1977, 1980). Frente al Sistema Evolutivo de Calificación de Koppitz (1983), el SCC muestra correlaciones más elevadas con criterios de rendimiento escolar en el estudio original (Brannigan & Brunner, 2002), así como en una muestra culturalmente diferente (en Hong Kong; Chan, 2002). El manual presenta una extensa revisión de los hallazgos psicométricos, así como los criterios de calificación de cada diseño; por ejemplo, los indicadores de consistencia interna y acuerdo interexaminadores son satisfactorios.

Procedimiento

El diseño de la investigación es post hoc, no experimental, y dentro de un marco cuyo objetivo es psicométrico, es decir, orientado hacia el instrumento de medición. Respecto al proceso de recolección de datos, este tuvo algunas diferencias en los dos grupos de edad. De este modo, a los niños del nivel 1 se les aplicó el TGB-modificado en una sesión de evaluación individual, junto a otras pruebas de desarrollo psicomotor como parte de una batería de evaluación de control del desarrollo, en que el TGB-modificado se administró generalmente al inicio de la sesión de evaluación. En los niños del nivel 2, se administró el TGB-modificado grupalmente, mediante cuadernillos en que cada figura se presentaba en una página distinta y en el tercio superior de la hoja orientada verticalmente. Este formato del TGB-modificado también se aplicó en el primer grupo de niños descrito. Para la administración en ambos grupos, se siguieron las reglas de aplicación estandarizada respecto al ambiente, relación con el niño e instrucciones generales de aplicación.

Por otro lado, en el proceso de calificación de los protocolos, tres estudiantes sirvieron como calificadores de los protocolos aplicados, en cada grupo de niños del nivel 1 y del nivel 2, pero que no participaron en la aplicación de las pruebas. El autor del presente estudio, con experiencia en evaluación psicológica a nivel profesional y de investigación, monitoreó el progreso de la administración, el entrenamiento en la calificación y el acuerdo entre los calificadores antes y después de concluido el estudio.

En el protocolo de entrenamiento la primera sesión sirvió para exponer el marco conceptual del SCC (Brannigan & Bruner, 2002) comparándolo con el Sistema Evolutivo de Koppitz (1984).

Seguidamente, se explicó el nivel de puntuación general y los criterios específicos para algunas de las láminas. En la segunda sesión, se pasó a calificar monitoreadamente al menos 5 de los ejemplos que aparecen en el manual, así como 5 protocolos de reproducciones hechas por niños; en tal sesión, se discutió la forma en que se llegó la calificación y se llegó a un acuerdo sobre la puntuación apropiada a cada uno de los 5 protocolos. En cada sesión de entrenamiento, se enfatizó la indicación clave del manual y del propio autor (G. Brannigan, comunicación personal, 2006), es decir: "que el dibujo debe ser tan bueno o igual que el que aparece en el manual; en caso de duda, se asignaría el puntaje más bajo". A cada calificadora se le asignó la tarea de calificar todos los protocolos, y luego pasar los protocolos a otra calificadora; se les instruyó para leer el manual y seguir estrictamente las indicaciones de calificación si hubiera dudas, y no consultar con las otras calificadoras.

Para obtener los resultados estadísticos respecto a la consistencia interna, se usó en coeficiente alfa (Cronbach, 1951), y luego se hicieron comparaciones de estas estimaciones en ambos grupos de edad, usando el programa ALPHATEST (Lautenschlager & Meade, 2008; Merino & Lautenschlager, 2003). Esta comparación permitirá revelar si estas estimaciones permanecen estables, cuando el impacto de las diferencias entre los puntajes de lo calificadores y entre ambos grupos de edad son factores causales de posible variación. Por otro lado, el análisis del acuerdo intercalificadores se condujo en dos niveles de puntajes: el puntaje total y los puntajes en los ítems. El puntaje total representa el nivel de desempeño en el niño sobre el atributo medido (integración visomotora), y su métrica se asume en el nivel de intervalo y continuo. Generalmente, los puntajes en el TGB son el objeto del análisis del acuerdo interexaminadores, pero nuestro estudio avanzó un paso más adelante, como se hizo en Fuller y Vance (1995), ya que evaluó el grado de acuerdo sobre cada ítem o lámina. El desempeño del niño en cada lámina está representado por una puntuación del 0 al 5 basado en el Sistema de Calificación Cualitativa de Brannigan y Brunner (1986, 2002); la puntuación en cada diseño es ordinal, y es una gradiente de exactitud del diseño reproducido.

El coeficiente utilizado para estimar el acuerdo intercalificadores estuvo en el contexto del modelo de componentes de varianza, usando el análisis de varianza desde el que se deriva el coeficiente de correlación intraclase (ICC) (Shrout & Fleiss, 1979), que se aplica cuando los datos bajo análisis tienen una métrica continua. Se aplicará el modelo 2, que asume que los calificadores son seleccionados aleatoriamente de alguna población de calificadores potenciales y cada calificador evalúa a cada examinado; este es el modelo de efectos aleatorios de dos vías y cubre mayormente las situaciones de acuerdo intercalificadores. Se calculará el ICC para estimar el acuerdo sobre una sola medición o calificador (ICC [2,1]). Debido que existe una correspondencia conceptual y algebraica entre el estadístico Kappa para múltiples calificadores sobre variables ordinales y los coeficientes de correlación intraclase (Fleiss & Cohen, 1973; Rae, 1988), se aplicó el ICC también a los puntajes individuales (seis láminas).

En la determinación del grado de acuerdo, los niveles cualitativos de acuerdo recomendados tienden a variar de autor en autor (Charter, 2003); pero se usará uno que es posiblemente de los más citados en la literatura: Cicchetti y Sparrow (1981) y Ciccheti (1994) que declaran 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. También se informa de la correlación de Pearson entre los examinadores, pero la interpretación del grado de acuerdo no se pondera de manera importante con este estadístico, ya que este únicamente informa de la relación lineal entre las variables y no considera en su estimación las diferencias entre los puntajes de los examinadores. Las correlaciones producto-momento, en esencia, son insensibles a la escala usada en los puntajes, pero sí a la monoticidad de las relaciones (Cone, 1999); por lo tanto, ya que nuestro interés es la variabilidad de la magnitud del acuerdo, la correlación intraclase será la más apropiada. Los análisis finalizan con la presentación de resultados descriptivos que re-

presentan los primeros datos de tendencia central y variabilidad en participantes hispanos, de este nuevo sistema de calificación.

Resultados

Presentamos a continuación, los análisis de la confiabilidad respecto a la consistencia interna y al acuerdo intercalificadores del Sistema de Calificación Cualitativo aplicado a la versión modificada del Test Gestáltico Visomotor.

Confiabilidad

Consistencia interna. En la Tabla 1 se muestra que la consistencia interna tendió a ser menor en los puntajes de las calificadores en el grupo del nivel de edad 2 ($X_{\alpha.Conbach} = 0.55$) respecto al otro grupo ($X_{\alpha.Conbach} = 0.78$). Dentro del nivel de edad 1, la consistencia interna tuvo pocas variaciones, pues la diferencia entre ellas no fue estadísticamente significativa, $\chi^2(2) = 3.07, p = 0.21$; en cambio, en el grupo de niños mayores (nivel 2), se detectaron diferencias estadísticamente significativas, $\chi^2(2) = 7.28, p = 0.02$. En este grupo, una calificadora C bajó la consistencia del grupo cuando fue pareada con las demás, y la variabilidad de sus puntajes estuvo relacionada con esta situación, ya que la desviación estándar de sus puntuaciones totales fue menor en ambos grupos de edad.

Acuerdo intercalificadores. Nuestros resultados señalan un acuerdo intercalificadores más elevado en el puntaje total y en cada diseño reproducido (Tabla 2) para el nivel de edad 1. Excepto dos láminas (diseño 2 y 3), el resto tuvo un patrón de acuerdo similar entre los dos niveles de edad; es decir, hubo diseños que aparentemente tienden a generar más acuerdo que otras. Cuantitativamente, el pobre nivel de acuerdo en las láminas 2 y 3 tuvo más influencia para explicar el más bajo nivel de acuerdo hallado en el puntaje total de los niños en el nivel 2; pero los demás ítems también revelaron que el acuerdo en ellos fue sistemáticamente menor que en el acuerdo de los niños del nivel 1. La observación de la Tabla 2 hace reconocer que el máximo nivel de acuerdo en el grupo del nivel 2 fue *bueno*, mientras que en el grupo del nivel 1, se alcanza hasta un nivel de acuerdo *excelente*. La Tabla 2 también muestra los resultados de aplicar el método de Alsawalmeh y Feldt (1992) con un programa ad hoc (Merino, en revisión) para comparar correlaciones intraclase de muestras independien-

Tabla 1
Estadísticos descriptivos, consistencia interna y correlaciones para los calificadores según los niveles

Calificadores	M	*DE*	Asimetría[a]	Curtosis[b]	.α Cronbach	*.r Pearson*		
						A	B	C
Nivel 1								
A	15.40	3.55	0.197	-0.490	0.75	1		
B	14.52	3.78	0.332	-0.215	0.81	0.91	1	
C	14.90	3.50	0.201	-0.656	0.78	0.88	0.93	1
Nivel 2								
A	24.41	2.25	-0.309	-0.065	0.63	1		
B	24.20	2.31	-0.343	0.060	0.62	0.85	1	
C	22.61	1.84	0.183	-0.238	0.41	0.78	0.72	1

[a]: error estándar en el nivel de edad 1 = 0.36; y en el nivel de edad 2 = 0.35
[b]: error estándar en el nivel de edad 1 = 0.0.71; y en el nivel de edad 2 = 0.0.70

Fuente: elaboración propia.

tes (Merino, 2009). Como se observa, las diferencias estadísticamente significativas ocurrieron en la mitad de las primeras láminas, así como en el puntaje total. Es aparente una secuencia curvilínea de diferencias entre las láminas.

Análisis descriptivos

Normalidad. Las pruebas de normalidad basadas en Komogorov-Smirnov (*KS* con ajuste Lilliefors; Lilliefors, 1967) y en Shapiro-Wilk (*SW*, Shapiro & Wilk, 1965) indicaron que no hay alejamientos sustanciales de la distribución normal teórica para el grupo de nivel 1 ($KS < 0.11$, $SW < 0.99$) y del nivel 2 ($KS < 0.15$, $SW < 0.97$). Esto sugiere que se podrían usar los percentiles de esta distribución teórica para describir las posiciones de rendimiento de los niños evaluados.

Variabilidad y tendencia central. Se observa que la calificadora C tuvo menor variabilidad que los puntajes totales de las otras calificadoras (Tabla 1); esta tasa de variabilidad fue de 5 % y 23 % mayor en los niveles de edad 1 y 2, respectivamente. Sin embargo, las diferencias entre las varianzas no fueron estadísticamente significativas en el nivel 1 ($W\ Mauchy = 0.93$, $\chi^2(2) = 2.51$) y nivel 2 ($W\ Mauchy = 0.91$, $\chi^2(2) = 3.74$).

Tabla 2
Correlaciones intraclase (ICC) para el puntaje total y puntajes asignados a cada lámina, para el nivel 1 y nivel 2 de edad

	ICC		
	Nivel 1	**Nivel 2**	**Prueba T (Alsawalmeh & Feldt, 1992)**
Puntaje total	0.89 [0.83 – 0.94] Excelente	0.64 [0.29 – 0.82] Bueno	F(61, 33) = 3.27**
Lámina 1	0.85 [0.76 – 0.91] Excelente	0.66 [0.52 – 0.78] Bueno	F(59, 35) = 2.26**
Lámina 2	0.74 [0.60 – 0.84] Bueno	0.29 [0.08 – 0.49] Pobre	F(86, 45) = 2.73**
Lámina 3	0.91 [0.86 – 0.95] Excelente	0.29 [0.11 – 0.48] Pobre	F(86, 36) = 7.88**
Lámina 4	0.76 [0.63 – 0.85] Excelente	0.71 [0.56 – 0.82] Bueno	F(55, 38) = 1.20
Lámina 5	0.72 [0.59 – 0.83] Bueno	0.60 [0.44 – 0.74] Bueno	F(62, 42) = 1.42
Lámina 6	0.62 [0.44 – 0.77] Bueno	0.74 [0.61 – 0.86] Bueno	F(52, 46) = 1.46

**: $p < 0.01$.

Fuente: elaboración propia.

Respecto a las diferencias en los puntajes promedio, en el nivel de edad 1, únicamente las diferencias entre el calificador A y B fueron estadísticamente significativas, $t(41) = 3.68$, $p < 0.01$; la magnitud del efecto para muestras relacionadas (Lipsey & Wilson, 2000), fue *d* de Cohen = 0.57. Mientras, en el nivel 2, las diferencias estadísticamente significativas surgieron entre los pares de jueces A – C ($t[43] = 8.56$, $p < 0.001$, *d* de Cohen = 1.78) y B – C ($t[43] = 6.67$, $p < 0.001$, *d* de Cohen = 1.01.

Las diferencias en los puntajes promedio entre los grupos de edad preescolar y escolar (nivel 1 y 2, respectivamente) fueron esperables debido a la maduración de sus funciones visomotoras, y que en estas edades tiende a incrementarse aceleradamente (Decker, 2008). Promediando los estadísticos descriptivos (media y desviación estándar de las calificadoras), el grupo de mayor edad superó el desempeño visomotor más de dos desviaciones estándares (*d* Cohen = 2.95).

Discusión

La evaluación de la integración visomotora continua siendo importante y predictiva (Sattler, 2003; Simner, 1991; Simner & Barnes, 1991) y sería raro que deje de incluirse en baterías predictivas para, por ejemplo, el inicio del primer grado (Berdicewski & Milicic, 2004). Por lo tanto, la evaluación las propiedades psicométricas de recientes propuestas de su medición, específicamente en el mundo hispano, es necesario para establecer científicamente su precisión para obtener interpretaciones adecuadas; y este es la situación del Sistema Cualitativo de Calificación-SSC, (Brannigan & Brunner, 2002), para la Prueba Gestáltica de Bender Modificada. En esta línea, el presente artículo examinó el error de medición mediante una estrategia de estimación puntual de la consistencia interna y del acuerdo intercalificadores, y de comparación de estas estimaciones en dos grupos de edad y con un mismo grupo de calificadores. En la búsqueda de evidencias psicométricas de los instrumentos de medición, se examinó el impacto de varias fuentes de error, especialmente de las diferencias entre calificadores y la consistencia interna (Anastasi & Urbina, 1997); y si ocurre una relación entre el nivel de desempeño en los dibujos y el nivel de acuerdo entre los calificadores, tal como se ha reportado y sugerido en investigaciones similares (Aylward & Smidth, 1986; Reynolds & Hickman (2004).

En lo concerniente a la consistencia interna, se halló que las estimaciones fueron relativamente bajas entre los niños de más edad, y que estuvieron relacionadas con el grado de acuerdo entre los calificadores así como en la dispersión de los puntajes. Los diseños reproducidos por niños de mayor edad tendieron a ser menos variables, y dado que sus funciones integrativas son más desarrolladas en los niños de menor edad, sus puntajes tendrán distribuciones asimétricamente negativas y posiblemente menos variables. Este resultado debe sugerir que el entrenamiento así como una adecuada estimación del acuerdo, deben ser componentes importantes en el control de calidad de evaluaciones que demanda el juicio o un elevado grado de subjetividad en la asignación de los puntajes. La relativa inestabilidad de la consistencia interna en el presente estudio, aún no podría considerarse como un resultados fijo, pues el coeficiente alfa, así como otras estimaciones derivadas de la teoría clásica de los test son dependientes de la muestra (Feldt & Brennan, 1989) puede ser un efecto del tamaño muestral.

En lo concerniente al acuerdo intercalificadores, hemos hallado un adecuado consenso entre los puntajes totales que generan varios calificadores, pero, en contraste, se observaron discrepancias en el acuerdo a nivel del ítem. También se encontró que el acuerdo fue mejor en la calificación de los protocolos de los niños de menor edad. Es posible que los calificadores tendieran a usar los criterios de puntuación ilusoriamente más confiadas, ya que su experiencia y acierto durante la calificación de los protocolos de los niños de menor edad les dio más seguridad en el manejo del SCC, y recurrieran con menor frecuencia a los ejemplos de puntuación de manual. Por lo tanto, en la situación de dudar qué puntuación asignar a una reproducción, no las habrían comparado con los ejemplos del manual y habrían aplicado inconsistentemente el criterio

clave. Esta evaluación hecha por cada calificador podría generar las inconsistencias y mayor desacuerdo al calificar a los niños del segundo nivel de edad. Este problema tiene especial importancia para el entrenamiento en pruebas visomotoras, considerando que tienden a utilizarse con frecuencia en la evaluaciones neuropsicológicas y psicopedagógicas (Sattler, 1996). Los resultados sugieren que se hallarían discrepancias severas entre los calificadores si se examina la calificación a nivel del ítem, y la dificultad de la reproducción no es efecto fijo que causalmente puede explicar estas discrepancias. Esto no apoya la afirmación de Aylward y Smidth (1986) ni la hipótesis de Reynolds y Hickman (2004), pues en este estudio se halló el patrón opuesto: los dibujos de los niños de mayor edad fue el contexto del mayor desacuerdo entre los calificadores.

Pero el acuerdo hallado, en general, ha sido aceptable para uso clínico, y es comparable con lo reportado en el manual; sin embargo, en los estudios (como los reportados por el manual) generalmente se usan coeficientes de correlación de Pearson en lugar de otros más apropiados y que sean sensibles no solo a los cambios monotónicos de los puntajes, sino también a la magnitud de los mismos, según han sido asignados por los calificadores (Cone, 1999). Por lo tanto, nuestras estimaciones del acuerdo suponen una mejor estimación que lo reportado por el manual.

Una limitación de nuestro estudio es que la diferencia en el desempeño ocasionadas entre la aplicación individual y grupal, podría ser una hipótesis rival que podría explicar una parte de las diferencias entre los grupos de edad, pero las autoras consideran que son de menor impacto; hay referencias que la variación en la administración no produce efectos fijos, sino más bien aleatorios, que bajo condiciones estandarizadas, no alteran el desempeño visomotor (Caskey & Larson, 1977, 1980)

Los manuales de las pruebas reportan importante información, pero dado que generalmente se usa la teoría clásica de los test para la evaluación psicométrica, estos son dependientes de la muestra (AERA, APA & NCME, 1999); por lo tanto, deberían replicarse la confiabilidad en la muestra de estudio e incluso en la práctica profesional mediante la solicitud a otro profesional que califique también las reproducciones de los niños (Williams et al., 2006). Finalmente, una vez establecido el acuerdo entre calificadores del equipo de trabajo o investigación, se puede calificar el DAP:IQ independientemente y sus puntajes intercambiables con otros calificadores.

El diseño descriptivo de nuestro estudio no permite hacer inferencias causales (Christensen, 2001) sobre el motivo del desacuerdo o el acuerdo, y limita la eficiencia del control sobre numerosas fuentes que podrían invalidar nuestras conclusiones, pero puede dar un respaldo favorable a la generalizabilidad de los resultados obtenidos, considerando que las condiciones de recolección y evaluación del acuerdo, se pueden empatar con lo que usualmente el profesional encuentra en su práctica psicopedagógica, en la enseñanza de pruebas psicológicas o en la investigación de campo. Un análisis normativo no fue posible en este estudio, pues el tamaño muestral no permitiría obtener estadísticos estables (Nunally & Bernstein, 1995); por lo tanto, una evaluación parcial de la universalidad de este sistema como un indicador de desarrollo vasomotor, requerirá un mayor tamaño muestral y un adecuado diseño de muestreo para obtener normas representativas. Aún con estas limitaciones, la presente investigación exploró la normalidad de la distribución de los puntajes, y ésta se acercó a tal distribución como para hacer razonables ajustes para determinar normas provisionales. Sin embargo, una muestra de mayor tamaño y demográficamente representativa debe ser un prerrequisito para obtener normas confiables. Por lo tanto, los parámetros psicométricos hallados aquí, dan un aporte inicial para mirar en esta dirección cuando se busca por recursos modernos de evaluación del área visomotora.

Considerando el examen del acuerdo intercalificadores ejemplificado aquí, debe ser un análisis sine qua non en medidas que involucran la subjetividad del evaluador, y desde la cual se produce la principal fuente de error en los puntajes en este tipo de evaluaciones (Anastasi & Urbina, 1997); y se puede considerar válida esta recomendación

en situaciones de investigación como en la experiencia del evaluador en práctica profesional. Por último, respecto al sistema de calificación de Brannigan y Brunner, las evidencias aquí presentadas siguieren una potencial herramienta para la evaluación de niños en el área escolar. Dado que esta investigación es una iniciativa psicométrica no concluyente, se prescribe la continuidad de la investigación en esta línea, y la expansión del estudio hacia los correlatos de rendimiento, inteligencia, y aspectos emocionales que definan la red nomológica de esta versión del TGB.

Referencias

Alsawalmeh, Y. M. & Feldt, L .S. (1992). Test of the hypothesis that the intraclass reliability coefficient is the same for two measurement procedures. *Applied Psychological Measurement, 16,* 195-205.

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

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

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

Bender, L. (1987). *El Test Guestáltico Visomotor: usos y aplicaciones clínicas*. Buenos Aires: Paidós.

Berdicewski, O. & Milicic, N. (2004). *Prueba de Funciones Básicas* (35ta ed.). Santiago: Editorial Universitaria.

Brannigan, G. G. & Brunner, N. A. (1989). *The modified version of the Bender-Gestalt Test for preschool and primary school children.* Brandon, VT: Clinical Psychology.

Brannigan, G. G. & Brunner, N. A. (1996). *The modified version of Bender-Gestalt Test for preschool and primary school children* (Revised). Brandon, VT: Clinical Psychology Publishing.

Brannigan, G. G. & Brunner, N. A. (2002). *Guide to the Qualitative Scoring System for the modified version of the Bender-Gestalt Test* (2nd ed.). Chicago, IL: Charles C. Thomas.

Brannigan, G. G. & Decker, S. L. (2003). *Bender Visual-Motor Gestalt Test* (2nd ed.). Itasca, IL: Riverside Publishing.

Brannigan, G. G. & Decker, S. L. (2006). The Bender-Gestalt II. *American Journal of Orthopsychiatry, 76,* 10-12.

Caskey, W. E., Jr. & Larson G. L. (1977). Two modes of administration of the Bender Visual-Motor Gestalt Test to kindergarten children. *Perceptual and Motor Skills, 45*(3), 1003-1006.

Caskey, W. E., Jr. & Larson, G. (1980). Scores on group and individually administered Bender Gestalt Test and Otis Lennon IQs of kindergarten children. *Perceptual and Motor Skills, 50,* 387-390.

Chang, P. W. (2001). Comparison of visual motor development in Hong Kong and USA assessed on the Qualitative Scoring System for the Modified Bender Gestalt Test. *Psychology Reports, 88,* 236-240.

Chang, P. W. (2002). Relationship of the visual motor development and academic performance in young children in Hong Kong assessed in the Bender-Gestalt Test. *Perceptual and Motor Skills, 90,* 209-214.

Charter, R. A. (2003). A breakdown of reliability coefficients by test type and reliability method, and clinical implications of low reliability. *The Journal of General Psychology, 130*(3), 290-304.

Cicchetti, D. V. & Sparrow, S. S. (1981). Developing criteria for establishing the interrater reliability of specific items in a given inventory. *American Journal of Mental Deficiency, 86,* 127-137.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and estandardized 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). New York: John Wiley & Sons.

Christensen, L. B. (2001). *Experimental methodology* (8th ed.). Needham Heights, Massachusetts: Allyn & Bacon.

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

Decker, S. L. (2008). Measuring growth and decline in visual-motor processes with the Bender Gestalt second edition. *Journal of Psychoeducational Assessment*, *26*(1), 3-15.

Feldt, L. S. & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105-146). New York: American Council on Education/Macmillan.

Fuller, G. B. & Vance, B. (1995). Interscorer reliability of the modified version of the Bender-Gestalt Test for preschool and primary school children. *Psychology in the Schools, 32*(4), 264-266.

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.

Jansky, J. & de Hirsch, K. (1972). *Preventing reading failure: Prediction, diagnosis, intervention*. New York: Harper and Row.

Johnston, C. W. & Lanak, B. (1985) Comparison of the Koppitz and Watkins Scoring Systems for the Bender Gestalt Test. *Journal of Learning Disabilities, 18*(7) 377-378.

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

Köppitz, E. M. (1984). *El test gestáltico visomotor para niños* (10ª ed.). Buenos Aires: Guadalupe.

Lautenschlager, G. J. & Meade, A. W. (2008). Alpha-Test: A windows program for tests of hypotheses about coefficient alpha. *Applied Psychological Measurement, 23,* 502-503.

Lipsey, M. W. & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks: Sage.

Lilliefors, H. W. (1967). On the Kolmogorov-Smirnov Test for Normality with mean and variance unknown. *Journal of the American Statistical Association*, *62,* 399-402.

Merino, C. & Lautenschlager, G. (2003). Comparación estadística de la confiabilidad alfa de Cronbach: aplicaciones en la medición educacional. *Revista de Psicología, 12*(1), 129-139.

Merino, C. (2009). *ICC COMPARE: A MS Excel program for testing equality of intraclass reliability coefficients.* Documento no publicado.

Morsbach, G., Priori, C. D. & Firnell, J. (1975). Two aspects of scorer reliability in the Bender-Gestalt test. *Journal of Clinical Psychology, 31*(1), 90-93.

Nunally, J. & Bernstein, J. (1995). *Teoría psicométrica*. México, DF: McGraw-Hill.

Palisano, R. L. & Dichter, C. G. (1989). Comparison of two tests of visual-motor development used to assess children with learning disabilities. *Perceptual and Motor Skills, 68*(3), 1009-1103.

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

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

Reynolds, C. R. & Hickman, J. A. (2004). *Draw-A-Person Intellectual Ability Test for children, adolescents, and adults (DAP:IQ)*. Austin: PRO-ED.

Reynolds, C. R. (2007). *Koppitz Developmental Scoring System for the Bender Gestalt Test: Examiner's manual* (2nd ed.). Austin, TX: Pro-Ed.

Rossini, E. D. (1993). The Bender-Gestalt psychopathology scale: Failure to infer validity in a school-aged sample. *Journal of Personality Assessment, 60*(3), 605.

Shapiro, S. S. & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika, 52,* 591-611.

Sattler, J. (2003). *Evaluación de la inteligencia infantil y habilidades especiales*. México: Manual Moderno.

Simner, M. L. (1991). Estimating a child's learning potential from form errors in a child's printing. In J. Wann, A. M. Wing & N. Sovik (Eds.), *Development of graphics skills: Research, perspectives, and educational implications* (pp. 205-222). London: Academic Press Inc.

Simner, M. L. & Barnes, M. J. (1991). Relationship between first-grade marks and the high school dropout problem. *Journal of School Psychology, 29,* 331-335.

Sugar, F. R. (1995). *Sugar Scoring System for the Bender-Gestalt Test*. Boston, MA: Educator Publishing Service.

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.

Watkins, E. O. (1976). *Watkins Bender-Gestalt Scoring System*. Novato, CA: Academic Therapy Publications.

Williams, T. O., Jr., Fall, A., Eaves, R. C. & Woods-Groves, S. (2006). The reliability of scores for the Draw-A-Person intellectual ability test for children, adolescents, and adults. *Journal of Psychoeducational Assessment, 24*(2), 137-144.