EXPLORACIÓN DE DIFERENCIAS NORMATIVAS EN EL SISTEMA DE CALIFICACIÓN CUALITATIVA PARA EL TEST GESTÁLTICO DE BENDER MODIFICADO

EXPLORING NORMATIVE DIFFERENCES IN QUALITATIVE SCORING SYSTEM FOR MODIFIED THE BENDER GESTALT TEST

César Merino Soto*
Universidad de San Martín de Porres, Perú.

Recibido: 22 de enero de 2011 Aceptado: 12 de setiembre de 2011

RESUMEN

El presente estudio explora la magnitud de las diferencias en los puntajes del Sistema de Calificación Cualitativa para el Test Gestáltico de Bender Modificado, usando diferente información normativa proveniente de Perú, Estados Unidos y China. En una muestra de 324 niños(as) peruanos entre 5 y 6 años de edad, se analizaron las potenciales diferencias en la densidad, tendencia central, dispersión y clasificaciones de rendimiento visomotor. Se hallaron grandes diferencias normativas, y por lo tanto, el desempeño en los participantes se vio altamente sobreestimado o subestimado dependiendo de la norma usada. Se discute el impacto de estos resultados en la apropiada práctica evaluativa en niños.

Palabras clave: Evaluación, intercultural, Test de Bender, sistema calificación cualitativo.

ABSTRACT

This study explores the magnitude of difference in scores from Qualitative Scoring System to Bender Gestalt Test Modified using different normative data from Peru, USA and China. In a sample of 324 children (boys and girls) between 5 and 6 ages, we analyzed the potential differences in density, central tendency, dispersion and visual motor performance classifications. It was found large normative differences, and therefore, performance in participants was highly overestimated or underestimated depending on the standard used. It discusses the impact of these results in the appropriate assessment practice in children.

Key words: Assessment, Bender Gestalt Test, intercultural, Qualitative Scoring System

El *Test Gestáltico de Bender* (TGB; Bender, 1938) continúa siendo, desde hace varias décadas, una de las pruebas más populares, y su uso ha generado más de mil artículos de investigación (Brannigan & Decker, 2003). Hasta la fecha, se observan muchos sistemas de calificación para el TGB, y cuya longevidad en la historia de la evaluación psicológica hace que generalmente el TGB no requiera mucha presentación respecto a su origen y popularidad internacional. En el conocimiento de este método de evaluación, se debe diferenciar entre el test propiamente dicho (estímulos, administración) y el sistema de calificación para puntuar las reproducciones. Justamente, uno de los sistemas más longevos y exitosos métodos es el *Sistema de Calificación Evolutiva de Koppitz* (*SCE*), publicado originalmente en Koppitz (1963), pero fue diez años después (Koppitz, 1975), en que se refinó y mejoró la información normativa para incluir grupos minoritarias y mayor representatividad muestral (Makhele, Walker & Esterhuyse, 2006).

Pero con el advenimiento en esta década de las modificaciones estructurales y funcionales importantes del TGB (por ejemplo, Brannigan & Brunner, 2002; Brannigan & Decker, 2003; Sisto, Noronha & Santos, 2006; Reynolds, 2007), se promovería un resurgimiento de las investigaciones relevantes para aportar datos psicométricos y comparaciones normativas. Estas modificaciones actuales señalan una amplitud en el escalamiento de los puntajes del TGB por medio de la adición de ítems fáciles y difíciles, así como en cambios en los procedimientos de puntuación

* sikayax@yahoo.com.ar

a los diseños reproducidos (Brannigan & Brunner, 2002; Brannigan & Decker, 2003; Reynolds, 2007; Sisto et al., 2006).

La potencial utilidad intercultural del TGB parece haber promovido los esfuerzos para evaluar las normas propuestas por Koppitz (1963; 1975), y generar las propias cuando la información paramétrica original de Koppitz no mostraba una apropiada «bondad de ajuste» con el contexto cultural de las investigaciones normativas. Junto con el examen de las normas originales, las investigaciones normativas con el sistema de Koppitz evaluaban y confirmaban también la hipótesis maduracional, que declaraba una covariación negativa entre la edad y el número y tipo de errores observados. Los nuevos datos hallados en diversas culturas (por ejemplo, Chang, 1990; Ghassemzadeh, 1988; Yousefi et al. 1992; Mazzeschi & Lis, 1999) aún respaldan esta hipótesis.

Por otro lado, algunos estudios con el TGB han mostrado diferencias más allá del error de muestreo entre los grupos que muestra diferentes grados de aculturización occidental. Por ejemplo, en muestras africanas (Karr, 1982; Makhele, Walker & Esterhuyse, 2006) y de medio oriente (Katz, Kizony & Parush, 2002; Parush, Sharoni, Hahn-Markowitz & Katz, 2000; Rosenblum, Katz, Hahn-Markowitz, Mazor-Karsenty & Parush, 2000) se hallaron diferencias frente a las normas americanas de Koppitz, (1963, 1975) así como diferencias intra-grupos (dentro de las mismas culturas). Estas discrepancias parecieron estar influenciadas por el monto de educación recibida y por el nivel de pobreza (por ejemplo, ver Rajabi, 2009). Generalmente, el hallazgo más frecuente es que, comparados con las normas americanas publicadas por Koppitz los países con menor influencia occidental y más desigualdad educativa y económica, los puntajes los puntajes tienden a ser menores, (por ejemplo, Karr, 1982; Makhele et al., 2006). Estos hallazgos convergen con las primeras revisiones de las diferencias culturales y socioeconómicas y educativas en grupos minoritarios dentro de los Estados Unidos (Buckley, 1978).

Una de las mayores distinciones entre la variedad de versiones y modificaciones del TGB es respecto al enfoque objetivo versus el enfoque global (Dana, Field, & Bolton, 1983); esta diferencia se refiere a los métodos de calificación de los diseños reproducidos, que respectivamente enfatizan evaluaciones con menor o mayor subjetividad en la obtención de los puntajes. Los tempranos aportes en la línea de los enfoque globales han provenido de Pauker (1976), Keogh y Smith (1961) y de los trabajos de deHirsh (deHirsch, Jansky & Langford, 1966; Jansky & deHirsch, 1972). En esta línea, uno de los métodos más recientes es el Sistema de Calificación *Cualitativa* (*SCC*, Brannigan & Brunner, 2002). Aunque recibió la influencia de Bender (1938), esencialmente se derivó de dos trabajos realizados en los años 60s (Keogh & Smith, 1961; deHirsch et al., 1966), quienes coordinaron con L. Bender para elaborar un sistema de puntuación simplificado que permitiera la evaluación global de los diseños reproducidos por los niños. El SCC es un refinamiento actual del sistema creado por Jansky y deHirsch (1972), y utiliza un puntaje para cada diseño que va desde 0 (dibujos aleatorios, garabatos y ausencia de diseño) hasta 6 (representación exacta del diseño). Este es un enfoque relativamente nuevo para calificar las reproducciones obtenidas de la presentación del Test Gestáltico de Bender Modificado, y hasta la fecha de elaboración del presente estudio, no se ha publicado alguna investigación fuera del Perú sobre este potencial método aplicable a niños preescolares. Aunque se han detectado diferencias normativas entre la información paramétrica en China (Chan, 2001) y la muestra norteamericana de estandarización del SCC (Brannigan & Brunner, 2002), no se han explorado hasta la fecha estas posibles discrepancias con otros grupos, como por ejemplo, en Latinoamérica.

Los aportes más frecuentes sobre las diferencias normativas provienen sin duda del uso del sistema de Koppitz, en que algunos de tales trabajos con población hispana han mostrado resultados inconsistentes. Por ejemplo, los niños españoles muestran un crecimiento más acelerado (Aguirre, Cortadellas & Tuset, 1988), mientras que los niños mexicanos y portorriqueños (Román & Vázquez, 1984) muestran un crecimiento menos acelerado, tal como ha sido replicado recientemente con muestras mexicanas (Fernández & Tuset, 2007). En Argentina, las normas de Casullo (1988) no diferían severamente de las primeras normas de Koppitz (1963), pero datos recientes muestran una superioridad frente a ambas normas (Pelorosso & Etchevers, 2004). En Brasil, esos patrones también se repiten (Kroeff, 1988, 1992). Tal como se ha expuesto desde los primeros trabajos normativos en otras culturas, la velocidad y pendiente de estos cambios generalmente

* sikayax@yahoo.com.ar

ocurren en los primeros años de edad, y mientras la edad avanza las diferencias interculturales se hacen menos intensas.

Exceptuando el trabajo de Chang (1990), en Perú, no ha habido trabajos publicados que puedan identificarse como representativos respecto a la comparación normativa entre muestras peruanas y normas existentes para alguna versión del TGB. La información normativa de Chang (1990) fue excepcional debido al tamaño muestral, pero no efectuó comparaciones estadísticas entre sus normas peruanas y las normas existentes para el sistema Koppitz (1963), así como otras evidencias de validez.

Por lo tanto, como parte de una línea de investigación que introduzca esta versión relativamente nueva del TGB en el habla hispana, y específicamente en Perú, el presente estudio evaluará las diferencias normativas en los puntajes de un grupo de niños peruanos, usando el SCC como método de puntuación para la versión modificada del TGB. Ni el Sistema de Calificación Cualitativa (SCC) de Brannigan y Brunner ni la forma abreviada del TGB han sido objeto de investigaciones normativas en Iberoamérica, excepto en Perú, donde hay una creciente información psicométrica publicada que señala buenas cualidades psicométricas (Merino, 2009, 2010a, 2010b; Merino & Benites, 2011).

Método

Participantes

La población del presente estudio son los niños(as) en etapa de ingreso al primer grado de primaria de colegios estatales de un distrito al sur de Lima Metropolitana, provenientes de instituciones educativas públicas y privadas dentro de la misma localidad. La población estudiada pertenece a la jurisdicción del UGEL 07 (Unidad de Gestión Educativa Local), según esta registrada en el Ministerio de Educación de Perú. Este distrito es considerado como una población menos pobre (quintil 4); sus indicadores de pobreza son menos severos que otros distritos de Lima Metropolitana (FONCODES, 2006).

La muestra de participantes (n = 324; varones = 196, 60.5%) provendrá de tres instituciones educativas públicas de nivel primario y una de preescolar de la misma localidad. En estas instituciones educativas, la enseñanza es unidocente y los profesores son de género femenino, y el tamaño de cada clase de primaria va desde 30 hasta 35 alumnos. Los alumnos ingresantes al primer grado en colegios públicos no son seleccionados, ya que la matrícula es libre y universal (Merino, Díaz, Zapata & Benites, 2006). Los niños participantes no estuvieron en programas especializados o ad hoc para el estímulos de habilidades visomotoras o de lenguaje, y recibieron en general el monto de instrucción de acuerdo a la tasa profesor-alumno. La distribución de la edad de nuestros participantes, parece representar generalmente el rango de edad en el periodo de ingreso al primer grado de primaria en Perú (Tabla 1).

Las características funcionales, estructurales y organizacionales de los colegios públicos de primaria en Perú tienden a ser similares, y por lo tanto podría asumirse una razonable equivalencia de los niños del presente estudio respecto a los niños del mismo sistema escolar público en zonas urbanas.

El nivel educativo alcanzado por el jefe de familia de los niños evaluados se distribuye de la siguiente manera: primaria= 17 (5.3%), secundaria= 198 (61.1%), estudios técnicos= 78 (24.1%), y universitaria= 17 (5.3%). En estas familias, las madres tienden a pasar más horas con el niño, pues se ocupan del hogar, y eventualmente realizan trabajos independientes; y mayoritariamente, las familias de los niños conviven con otros miembros familiares . Por lo tanto, los hogares de los niños lo integran generalmente más de tres miembros, con padres de condición civil de casados o convivientes, y pertenecientes a la clase media o menos.

Instrumento

Test Gestáltico Visomotor – Modificado (TGB – M). Esta versión usa únicamente seis de los diseños originales (A, 1, 2, 4, 6 y 8) para su aplicación el niños preescolares hasta los primeros grados del nivel primario (4.5 hasta 8.5 años), dado que tales diseños son los más apropiados para niños pequeños (Brannigan & Brunner, 2002). Esta versión incluye un sistema de puntuación del desempeño gráfico del niño sobre las seis láminas presentadas, denominado *Sistema de Calificación Cualitativa* (*SCC*, Brannigan & Brunner, 2002) de 6 puntos, desde una puntuación de 0 (líneas aleatorias, garabateo, sin concepto de los diseños) hasta 5 (representación exacta del diseño). Esta versión se califica por un método de inspección global, que refleja el grado de diferenciación y la gestalt de los diseños

* sikayax@yahoo.com.ar

reproducidos. La investigación sobre la confiabilidad interna, test-retest e inter-jueces, y validez dan soporte a su valor métrico, produciendo buena diferenciación del desempeño visomotor durante la evaluación psicopedagógica (Brannigan & Brunner, 2002). Frente al Sistema de Koppitz, el SCC muestra correlaciones más elevadas con criterios de rendimiento escolar (Brannigan & Brunner, 2002; Chan, 2001). El manual presenta una extensa revisión de los hallazgos psicométricos, así como los criterios de calificación de cada diseño. En el presente estudio, la consistencia interna de los puntajes del SCC en nuestra muestra fue $\alpha = 0.77$.

Procedimiento

La recolección de los datos se efectuó dentro de un proceso de evaluación de niños ingresantes a dos colegios públicos, entre Octubre (2009) y Marzo del 2010. A los niños se les administró grupalmente una batería de pruebas de lápiz-papel, que exploraban conocimientos y habilidades preacadémicas. Considerando que se ha reportado su equivalencia con la modalidad de administración individual (Koppitz, 1963; Buckley 1978; Tolor & Brannigan, 1980), se administró grupalmente la versión abreviada del TGB-M, entre 6 a 12 niños por grupo, y en orden balanceado respecto a las otros instrumentos. Durante la evaluación del TGB-M, se mantuvieron las condiciones estandarizadas recomendadas para maximizar la varianza relevante al constructo evaluado (Bracken, 2000; McCallin, 2006) y en concordancia con las directrices para el uso apropiado de pruebas (Hambleton, 1996; International Test Commission, 2000).

La calificación de los protocolos con el SCC se hizo por el autor de la investigación y tres estudiantes de pregrado pertenencientes al tercio superior; previo a la obtención de los puntajes, se entrenó, monitoreo y evaluó el acuerdo entre los cuatro calificadores, durante dos sesiones de entrenamiento; la correlación intraclase entre los calificadores (modelo de dos vía aleatorias), fue 0.84. Por otro lado, para la obtención de los puntajes estandarizados locales, los puntajes directos se transformaron no linealmente, con ajuste a la normalidad, en puntajes T. Estos puntajes se compararán con los puntajes T obtenidos de transformar los puntajes directos usando las normas americanas y chinas (Brannigan & Brunner, 2002; Chan, 2001) separadamente para los grupos de edad congruentes con el manual.

Los análisis estadísticos examinarán la propiedades cuantitativas susceptibles de mostrar variaciones normativas, y serán descritas en la presentación de cada resultado.

Resultados

Los resultados se presentarán en dos partes: la primera son las propiedades estadísticas de cada puntaje normativo y sus variaciones respecto a la edad (Tabla 1). La segunda parte (el análisis principal), compara los puntajes de los datos peruanos, americanos y chinos en cada una de sus propiedades estadísticas, y en cada nivel de edad (Tabla 2).

Comparaciones dentro de los grupos normativos. Al aplicar un ANOVA para comparar los puntajes directos promedio entre las tres categorías de edad obtenidas por la muestra de participantes (Tabla 1), no se detectaron diferencias estadísticamente significativas entre ellos (F[2, 321] = 2.16, $p = 0.11$). Luego, para verificar el incremento lineal monotónico del puntaje directo promedio entre las categorías de edad, se aplicó un contraste lineal a priori. Se halló un incremento lineal de los puntajes entre las edades, $L = 0.915$, $F(1, 321) = 4.30$, $p = 0.03$. Este resultado da soporte la validez de constructo del *SCC* en relación a la edad usando los puntajes directos.

Entre tanto, las diferencias en las medias de los puntajes T derivados de las normas americanas y chinas, mostraron patrones diferentes. Usando las normas americanas, hubo diferencias estadísticas entre las edades (F[2, 321] = 6.13, $p = 0.002$), y estas diferencias en los puntajes T promedio también representaron una tendencia lineal, pero esta vez de manera monotónicamente decreciente, $L = - 4.14$, $F(1, 321) = 7.15$, $p = 0.008$. Esto indica que los puntajes T promedio usando las normas americanas decrecían sustancialmente a medida que la edad aumentaba. En la Tabla 1 se puede observar esta tendencia, debajo del encabezado «Puntaje T – USA» para cada edad. Por otro lado, las normas T chinas no produjeron diferencias estadísticas entre las edades ($F(2, 321) = 1.80$, $p = 0.16$), ni alguna tendencia sustancial entres las medias; sin embargo, se puede reconocer una leve tendencia lineal decreciente.

Finalmente, la normalidad de los puntajes se examinó con las pruebas de Shapiro Wilk (1965) y Kolmogorov-Smirnov con corrección Lilliefors (1967), así como

* sikayax@yahoo.com.ar

estadísticos de simetría y curtosis (Tabla 1). Un nivel de edad exhibió distribuciones aproximadamente normales, pero el resto de las edades mostraron distorsiones en la simetría, curtosis o ambos, particularmente en la edad 5.6 a 5.11 años.

La información de la Tabla 1 también indica que la forma distribucional puede mostrar cambios importantes que son dependientes de la información normativa usada. Por ejemplo, las distorsiones en la simetría tienden a ser similares en distribuciones de los puntajes directos de los niños peruanos y en los puntajes T derivados de la norma americana; mientras que el uso de las normas chinas siempre muestra diferencias con las anteriores. Por otro lado, las distribuciones empíricas son más parecidas a la distribución normal teórica en la edades 5.0 – 5.5, y 6.0 – 6.5. En la Figura 1, se muestra gráficamente la forma de las distribuciones de los puntajes T.

Tabla 1
Información descriptiva de los puntajes T obtenidos manual (USA y China) y de los puntajes directos de los participantes

	Edad 5.0 – 5.5 (n = 51)			**Edad 5.6 – 5.11 (n= 169)**			**Edad 6.0 – 6.5 (n = 104)**		
	Puntajes	**Puntajes T**		**Puntajes**	**Puntajes T**		**Puntajes**	**Puntajes T**	
	Directos	**USA**	**China**	**Directos**	**USA**	**China**	**Directos**	**USA**	**China**
Media	18.745	62.37	49.05	19.550	61.66	47.85	20.03	49.05	43.87
DE	3.637	10.55	23.5	3.777	14.19	19.09	3.430	11.39	17.09
Mín. – Máx.	11 - 28	-	-	5 - 28	-	-	7 - 27	-	-
Simetría	0.11	0.14	0.03	-0.67	-0.69	-0.26	-0.42	-0.42	-0.01
(z_s)	(0.32)	(0.44)	(0.105)	(-3.59)	(-3.68)	(-1.39)	(-1.78)	(-1.80)	(-0.07)
Curtosis	0.14	0.07	-0.22	1.29	1.38	-0.23	1.06	1.06	-0.42
(z_c)	(0.20)	(0.11)	(-0.33)	(3.68)	(3.79)	(-0.62)	(2.21)	(2.30)	(-0.91)
Prueba Shapiro-Wilk	0.97	0.98	0.976	0.95**	0.96**	0.975**	0.96	0.96**	0.98
Prueba Kolmogorov-Smirnov (KS)	0.091	0.091	0.086	0.11	0.12**	0.11**	0.079	0.073	0.093*

**: $p < 0.01$

Comparaciones entre los puntajes normativos

Diferencias en la densidad. Para evaluar las distribuciones empíricas, se usarán gráficos kernel y la prueba Kolmogorov-Smirnov para dos muestras (*KS* , Smirnov, 1939), la cual hace un análisis bivariado de la hipótesis nula de igualdad de las distribuciones empíricas. Esta prueba estadística es sensible a las diferencias debidas a la locación, dispersión y/ o forma de la función acumulativa empírica de los puntajes de los grupos (Pett, 1997). En la Tabla 2 se muestra que las diferencias en la densidad ocurren en todos los niveles de edad, y esto es generalmente mayor entre los parámetros peruanos y americanos. Por medio de gráficos *kernel*, la Figura 1 presenta las diferencias distribucionales en cada nivel de edad. Se observa que el origen de las desigualdades distribucionales ocurre en la dispersión y en la locación, lo que originan diferencias de variada magnitud en la curtosis y simetría, tal como se analizó anteriormente (ver Tabla 1).

Diferencias en la tendencia central. En la Tabla 2, se observan los resultados de las comparaciones entre los grupos normativos dentro de cada grupo de edad, respecto a los valores promedio. Se aplicó la prueba *t de Student* para muestras dependientes, pero debido a la asimetría distribucional de los puntajes, se usó el procedimiento *bootstrap* para obtener los valores *p* para evaluar la significancia estadística. Para atenuar el efecto de la no-normalidad distribucional, también se derivaron versiones robustas (Algina, Keselman & Penfield, 2005; Hogarty & Kromrey, 1999) de la estimación de la magnitud del efecto basada en diferencias estandarizadas, *d de Cohen* (Coe & Merino, 2003), mediante un macro ad hoc para el programa SAS (Kromrey & Coughlin, 2007) que usará las medias recortadas en el cálculo (d Cohen$_{trim}$). Los análisis muestran que, excepto en la comparación Perú-China en el primer y tercer nivel de edad, el resto de las diferencias *t* de Student fueron estadísticamente significativas. Las diferencias entre

* sikayax@yahoo.com.ar

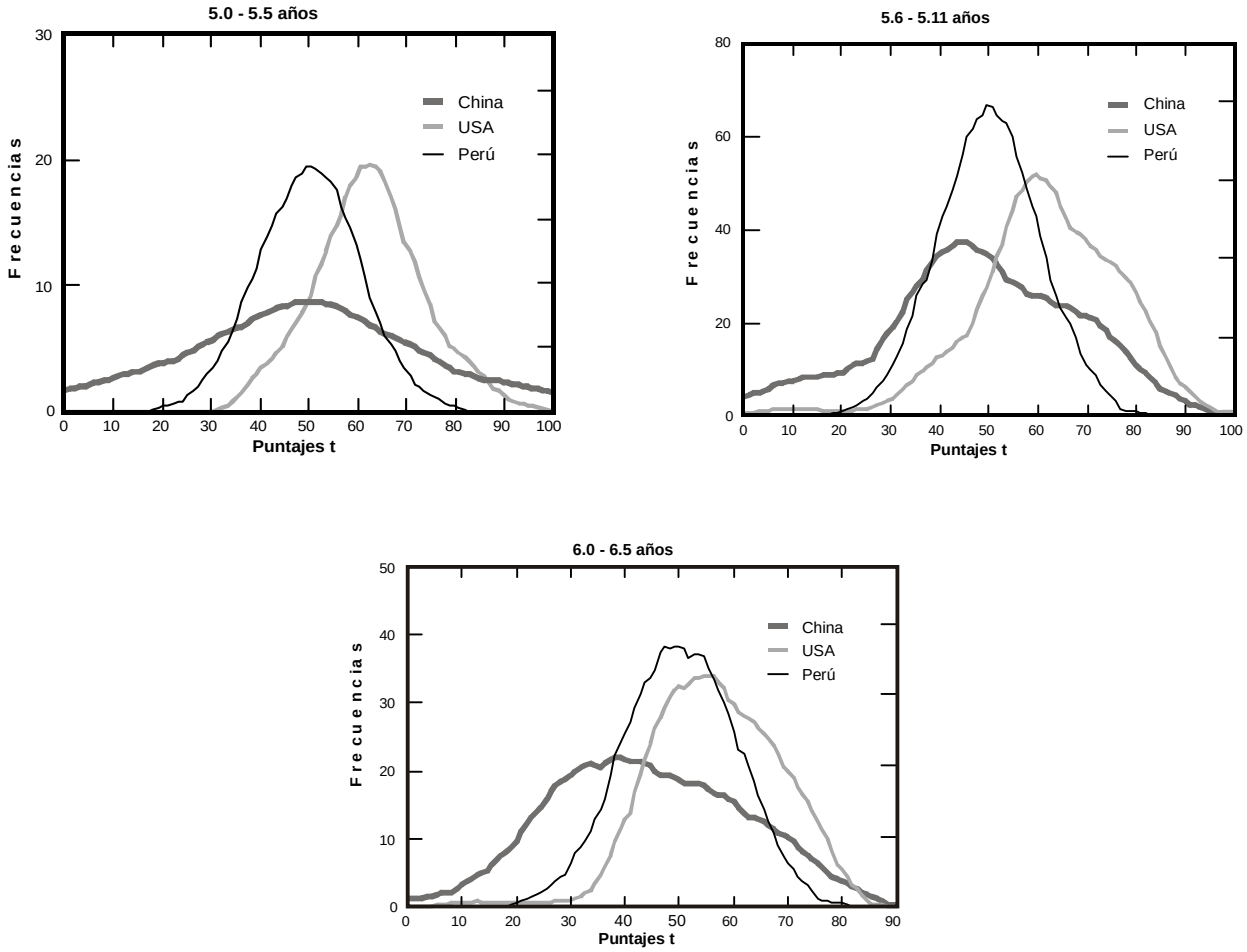

Figura 1. Comparación entre la distribución de los puntajes usando diferente información normativa

las normas peruanas y las normas americanas tendieron a disminuir conforme el incremento de la edad; y la magnitud de estas diferencias fue muy elevada, pues fueron mayores a 3 desviaciones estándares de las diferencias entre las medias. Las diferencias entre los promedios entre Perú y China aumentaron también con la edad, siendo más notorio en el tercer nivel de edad.

Diferencias en la varianza. Se compararon las varianzas entre cada par de DE mediante la una prueba de diferencias para varianzas relacionadas (Pitman, 1939; Howell, 1997). La variabilidad de los puntajes T extraídos de las normas americanas y chinas mostró diferencias estadísticamente significativas respecto a la variabilidad de los puntajes basados en los propios estadísticos de la muestra peruana. Las DE en puntajes T para la muestra peruana en los tres niveles de edad fueron 9.69, 9.86 y 9.81, respectivamente. Las diferencias en la variabilidad de los puntajes, en general, fueron más grandes entre la muestra peruana y las normas chinas, y la tendencia de estas diferencias fue que decrecieron con la edad, pero siempre la diferencia fue casi dos veces la magnitud de la dispersión de las normas chinas. Por ejemplo, de acuerdo a la razón de la dispersión de los puntajes peruanos (DE = 9.96) en la edad 5.0 a 5.5 frente la dispersión de la muestra china (DE = 23.5, Tabla 1), esta fue casi la mitad de esta última; en otras palabras, dos veces la dispersión de las normas peruanas. Por otro lado, la dispersión de las normas americanas frente a las normas peruanas fue menor, pero aún estadísticamente significativas según la prueba T de Pittman (Pitman, 1939; Howell, 1997).

* sikayax@yahoo.com.ar

Tabla 2
Comparaciones entre los puntajes empíricos y de los normas americanas y chinas derivados del manual (Brannigan y Brunner, 2002)

	Edad 5.0 – 5.5 (n = 51)		Edad 5.6 – 5.11 (n= 169)		Edad 6.0 – 6.5 (n = 104)	
	Perú vs USA	Perú vs China	Perú vs USA	Perú vs China	Perú vs USA	Perú vs China
A. Diferencias						
1. Distribución (KS)[a]	0.49**	0.25	0.46**	0.23**	0.26**	0.33**
2. Tendencia central						
t de Student	-70.82**	0.503	-31.28**	2.87**	-28.01	8.60**
Dif. Media	-12.33	0.98	-11.64	2.16	-6.44	6.21
[I.C. 95%]	[-12.67, -12.0]	[-2.75, 4.90]	[-12.32, -10.88]	2.16[0.71, 3.56]	[-6.79, -5.96]	[4.78, 7.67]
d Cohen$_{trim}$	-9.08	0.44	-5.57	0.50	-3.37	3.28
3. Variabilidad						
Ratio	1.089	2.42	1.43	1.93	1.16	1.74
T Pittman	3.34**	29.28**	12.86**	18.48**	4.68**	36.67**
B. Acuerdo						
ICC (2, 2)[b]	0.57	0.70	0.63	0.78	0.82	0.78
% de acuerdo	1.96%	27.45%	22.48%	35.50%	44.23%	29.80%
Kappa	-0.13	0.11*	0.043	0.18**	0.19**	0.16**
.*r* Pearson	0.99	0.99	0.98	0.97	0.98	0.99

**: p < 0.01. [a]. KS: Prueba Kolmogorov-Smirnov. [b]: ICC (2, 2) :correlación intraclase, dos vías aleatorias

Diferencias en las clasificaciones de desempeño. Para evaluar la concordancia de la estimación del desempeño usando normas distintas, se transformaron los puntajes T en clasificaciones cualitativas de rendimiento. Para obtener estas clasificaciones de rendimiento en niveles cualitativos, se categorizó el desempeño de los niños de acuerdo a las diferentes normas usadas. Se eligió clasificar el desempeño en 7 niveles de puntajes T: menos de 29, 30 hasta 35, 36 hasta 42, 43 hasta 55, 56 hasta 62, 63 hasta 69 y más de 70. Esta separación es consistente con las formas comunes de dividir el rendimiento en varias secciones para diferenciar mejor el rendimiento entre 2 desviaciones estándares alrededor de la media (Reynolds, 2007).

En todas las edades, el porcentaje de acuerdo en los puntajes clasificados según lo descrito en el párrafo anterior, estuvo por debajo del 50%, y mayormente alrededor del 30% de acuerdo. El acuerdo más pobre ocurrió en el primer nivel de edad, que prácticamente no se diferenció de cero. Estas estimaciones del acuerdo pueden estar sesgados por el acuerdo aleatorio, así que se usó el coeficiente Kappa (Cohen, 1960) para hacer una mejor estimación de la concordancia. Los coeficientes Kappa revelaron un peor desempeño clasificatorio entre las normas peruanas y no-peruanas. Aunque este acuerdo tendió a elevarse con el nivel de edad, los niveles de acuerdo hallados se consideran convencionalmente como *pobres* (Cicchetti, 1994). Por otro lado, usando el modelo de dos vías para las correlaciones intraclase (*ICC*; Shrout & Fleiss, 1979) y como límite mínimo el valor de 0.70 para separar a las calificaciones con aceptable precisión (Cicchetti, 1994), los resultados tendieron a ser algo bajos. El acuerdo, sin embargo, tendió a elevarse en los niveles de edad tardíos, y la magnitud de los *ICC* en el último nivel de edad podría considerarse satisfactorio considerando que superará al parámetro de comparación.

Discusión

La presente investigación se orientó a evaluar la información normativa de un nuevo sistema de calificación para el TGB-M, en niños peruanos. Los resultados hallados señalan que el uso de normas extranjeras respecto al *SCC* para el TGB-M ha exhibido un grado de variabilidad

* sikayax@yahoo.com.ar

paramétrica que compromete seriamente la práctica para aplicarlo en las decisiones educativas y administrativas en los niños hispanos, específicamente en Perú. Aunque se requiere un tamaño muestral mayor para establecer estadísticos más estables para compararlos frente las normas existentes de USA y China, tenemos la evidencia cuantitativa y contrastable para afirmar que las diferencias no son error de muestreo y que van más allá de la simple diferencia de medias. Este estudio aporta datos para una de las nuevas versiones del TGB aplicable a niños y que puede ser muy útil en una batería de evaluación del desarrollo o de despistaje de futuros problemas en el rendimiento escolar.

Ya que no hay otra información normativa disponible hasta la fecha sobre el uso de SCC en muestras hispanas a nivel mundial, no se puede conocer el grado en que la información normativa peruana presentada aquí puede diferir de otros grupos hispanos habitando en su propio país de origen, o grupos de hispanos inmigrantes USA.

Aunque la elección de la muestra del presente estudio es una limitación para generalizar nuestros resultados, se puede considerar que la magnitud de la baja probabilidad de cometer un error de Tipo I hallada en nuestra investigación, así como la magnitud de las diferencias en varias características estadísticas (la distribución, la tendencia central, la variabilidad y el acuerdo en la clasificación del rendimiento), han sido lo suficientemente grandes como para asumir que estas diferencias pueden ser consecuencia de influencias sistemáticas. Estas posibles influencias pueden ocurrir en las experiencias sociales y educativas de los niños participantes, lo cual es una explicación plausible y coherente con otras investigaciones que han hallado diferencias entre grupos culturales (Chang, 2001; Fernández & Tuset, 2007; Makhele et al., 2006). Las diferencias halladas no pueden ser ignoradas en las decisiones sobre la elección de las normas más apropiadas en el ejercicio del uso del *SCC*, e incluso en el uso de otras pruebas de desempeño visomotor. Como otros aspectos del desarrollo del niño, el impacto de la cultura depende de la calidad y cantidad de estímulos recibidos, así como de las interacciones en que estas ocurren, y debe considerarse una perspectiva emic-etic cuando se aborda el impacto diferencias de estas condiciones sobre la validez de un constructo (Kagitcibasi, 2004).

Uno de los resultados fue que los puntajes T promedio usando las normas americanas sistemáticamente sobreestimaron el desempeño en el BGT en los dos primeros niveles de edad, y se acercaron al parámetro central de 50 en el último nivel de edad. Esta sobreestimación significó que en el primer semestre de la edad de 6 años, la distorsión sea 3 desviaciones estándares debajo de la media, mientras que en todo el rango de la edad 5, la sobreestimación varió entre 9 y 5 unidades estandarizadas T. En todas las edades, la discrepancia entre los niños de la norma americana y los niños peruanos de nuestro estudio favorece claramente a estos últimos. Lo contrario ocurrió usando las normas chinas, en que los puntajes T tienden a subestimar el desempeño visomotor; esto ocurrió en los dos últimos niveles de edad y ligeramente debajo de 50 en el primer nivel de edad. Esto parece indicar que las diferencias entre el desempeño visomotor de niños chinos y los niños peruanos participantes se acentúan más cuando se acercan a los 6 años de edad, y que las discrepancias generalmente favorecen a los niños chinos en el rango de edad evaluado, pues el uso de las normas chinas producen una consistente subestimación de la habilidad visomotora en los niños peruanos de nuestra muestra.

Además que una de las limitaciones del estudio es el tamaño muestral de algunos rangos de edad evaluados, se tiene que la muestra solo proviene de un distrito urbano en Lima Metropolitana. Hay una pregunta abierta sobre la generalizabilidad de los presentes resultados, pero que la futura investigación podrá responder sobre este punto.

Nuestros resultados apuntan a concluir varios aspectos. En cada grupo de edad, la distorsión descriptiva es mucho mayor usando las normas americanas que usando las normas chinas; pero en ambas situaciones, la elección de cualquiera de estas normas es inapropiada porque ubican normativamente al niño evaluado en posiciones severamente equivocadas. En donde la diferencia en la locación fue menor y correspondiente a una moderada discrepancia (por ejemplo en los dos primeros niveles de edad en la comparación Perú-China), podría parecer que usar la norma extranjera podría llevar a solo leves o moderados distorsiones en la descripción del evaluado. Pero el impacto puede ser realmente severo si lo consideramos en el rango completo de los puntajes posibles, pues en alguno de los niveles de puntuación las diferencias pueden ser mayores que las que ocurren en otros sectores de la escala de puntajes. Por ejemplo, el efecto de la gran disimilaridad en la dispersión

* sikayax@yahoo.com.ar

supone que los puntajes más alejados de la media tenderán a ser altamente sobre-estimados o subestimados. Y una consecuencia práctica importante de esta distorsión es la generación de falsos positivos o falsos negativos, respecto a la identificación de habilidades relacionadas con problemas en el aprendizaje. La información que se presenta en el estudio no solo indica que hay diferencias estadísticamente significativas entre los grupos, sino también una estimación de la magnitud de estas diferencias. En los estudios revisados en la introducción de este artículo, no se reportan estimaciones de este tipo, y no se podría evaluar si las diferencias tienen también un valor práctico y no solo estadístico. Si el lector se pregunta cuánto podría ser la diferencia entre el puntaje obtenido por un niño en la versión modificada el Bender, usando normas propias, la norma americana o china, entonces con seguridad podríamos decir que habría una severa diferencia, y una consecuente distorsión en el diagnóstico. Esta distorsión, sin embargo, variaría con la edad.

En la práctica, el uso de normas foráneas trae problemas no solo diagnósticos y descriptivos, sino también éticos. Si un grupo numeroso de niños, de diferentes edades, es evaluado con una prueba para la cual hay normas extranjeras disponibles, y la descripción estadística de los grupos se hará con tales normas, lo que se puede anticipar es que habría una razonable incertidumbre sobre la apropiada descripción de los niños y grupos evaluados. Es decir, los niños podrían tener puntajes que signifiquen una sobreestimación, subestimación o una apropiada estimación de su desempeño. Si a este problema se añade el efecto del error de medición, entonces tendremos una estimación del nivel de habilidad que no sería útil en ninguna circunstancia aplicada. Las guías éticas sobre un uso responsable e informado de las pruebas desde varias fuentes (American Educational Research Association, American Psychological Association & National Council on Measurement in Education, 1999; International Test Commission, 2000) exigen el cumplimiento de las cualidades psicométricas de los instrumentos de medición, pero enfatizan más aún la responsabilidad sobre las decisiones en la elección de estos instrumentos y de las consecuencias de las inferencias derivadas de sus usos. Y usar normas foráneas sin una evaluación de su adecuabilidad en el contexto de su uso, no puede ser menos que un ejemplo de una práctica inapropiada.

El presente estudio corrobora lo problemático de esta situación en el terreno aplicado y de investigación, pues las estimaciones de desempeño obtenidas de normas inapropiadas crean una imagen inexacta del estatus conductual de un niño o del grupo evaluado. Pero desde un marco metodológico, esta investigación pudo revelar el impacto de las diferencias normativas en varias fuertes de información paramétrica. Por lo tanto, se puede considerar que un análisis expandido hacia tales fuentes puede dar una mejor figura de los cambios cuantitativos en las diferencias normativas de los instrumentos en estudio de adaptación psicométrica. Y por último, si las modificaciones al TGB son concurrentes con los desafíos clínicos y los métodos psicométricos que cada modificación produce (Dana, Field & Bolton, 1983), entonces la información normativa debe ser un espacio de evaluación con la mayor precisión posible. En tal situación, el estudio presentado aquí puede hacer una contribución normativa respecto a la nueva versión abreviada del TGB, sino también metodológica considerando los aspectos secuencialmente evaluados.

Referencias

Aguirre, G., Cortadellas, M. & Tuset, A. (1988). *Baremación del test de Bender*. Barcelona: Oikos-Tau.

Algina, J., Keselman, H. J. & Penfield, R. D. (2005). An alternative to Cohen's standardized mean difference effect size: A robust parameter and confidence interval in the two independent groups case. *Psychological Methods*, *10*(3), 317-328.

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.

Bender, L. (1938). A visual-motor gestalt test and its clinical use (Research Monograph Nro. 3). New York: *American Orthopsychiatric Association*.

Bracken, B. A. (2000). Maximizing content-relevant variance: The assessment situation. In B. A. Bracken (Ed.) *Psychoeducational assessment of preschool children, 3th. ed.* (pp. 33-44). Needham Heights, MA: Allyn & Bacon.

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

Brannigan, G. G., Aabye, S. M., Baker, L. A. & Ryan, T. G. (1995). Further validation of the Qualitative Scoring System for the modified bender-gestalt test. *Psychology in the Schools, 32*(1), 24-26.

* sikayax@yahoo.com.ar

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

Buckley, P. D. (1978). The Bender Gestalt Test: A review of reported research with school-age subjects, 1966-1977. *Psychology in the Schools, 15*(3), 327-338.

Casullo, M. M. (1988). *Test de Bender Infantil. Normas Regionales Argentinas*. Buenos Aires: Guadalupe.

Chan, 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. *Psychological Reports, 88*, 236-240.

Chang, G. (1990). *Nueva Escala de Maduración del Bender Infantil*. Lima: Biblioteca Andina de Psicología.

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

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

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37-46.

Dana, R. H., Field, K. & Bolton, B. (1983). Variations of the Bender-Gestalt Test: Implications for training and practice. *Journal of Personality, 47*(1), 76-84.

deHirsch, K., Jansky, J. J. & Langford, W. S. (1966). *Predicting reading failure*. New York: Harper and Row.

Fernaìndez, T. & Tuset, A. M. (2007). Bender performance and socioeconomic status in Mexican children: A cross-cultural study. *Perceptual and Motor Skills, 105*(3), 906-914.

FONCODES. (2006). *Focalización geográfica: Nuevo mapa de pobreza de FONCODES 2006*. Lima: Ministerio de la Mujer y Desarrollo Social.

Ghassemzadeh, H. (1988). A Pilot study of the Bender-Gestalt test in a sample of Iranian normal children. *Journal of Clinical Psychology, 44*, 787-792.

Hambleton, R. K. (1996). Adaptación de tests para su uso en diferentes idiomas y culturas: fuentes de error, posibles soluciones y directrices practicas. En J. Muñiz (Coord.), *Psicometría* (pp. 207-238). Madrid: Universitas.

Hogarty, K. Y. & Kromrey, J. D. (1999, August). *Traditional and Robust Effect Size Estimates: Power and Type I Error Control in Meta-Analytic Tests of Homogeneity*. Paper presented at the annual meeting of the American Statistical Association, Baltimore.

Howell, D. C. (1997). *Statistical methods for psychology* (4th ed.). Belmont, CA: Duxbury.

International Test Commission (ITC). (2000). *Guidelines on Test Use: Spanish Version*. Translation authorised by the Colegio Oficial de Psicólogos. ITC: Author.

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

Kagitcibasi, C. (2004). Culture and child development. In C. Spielberger (Ed.), *Encyclopedia of Applied Psychology*. (Vol I., pp. 329-338). Oxford, UK: Elsevier.

Karr, S. K. (1982). Bender Gestalt performance of Sierra Leone West African children from four subcultures. *Perceptual and Motor Skills, 55*, 123–127.

Katz, N., Kizony, R. & Parush, S. (2002). Visuomotor organization and thinking operations performance of school-age Ethiopian, Bedouin, and mainstream Israeli children. *Occupational Therapy Journal of Research, 22*, 34–43.

Keogh, B. K. & Smith, C. E. (1961). Group techniques and proposed scoring system for the Bender-Gestalt Test with children. *Journal of Clinical Psychology, 17*, 172-175.

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

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

Kroeff, P. (1988). Normas brasileiras para o Teste de Bender. *Psicologia: Reflexão e Crítica, 1/2* (3), 12-19.

Kroeff, P. (1992). Desempenho de crianças no Teste de Bender e nívelsócioeconômico-cultutal. *Psicologia: Reflexão e Crítica, 5* (2), 119-126.

Kromrey, J. D. & Coughlin, K. B. (2007). ROBUST_ES: A SAS® macro for computing robust estimates of effect size. *Proceedings of the Southeast SAS Users Group*.

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

Makhele, L., Walker, S. & Esterhuyse, K. (2006). Utility of the Koppitz norms for the Bender Gestalt Test performance of a group of Sesotho-speaking children. *Journal of Child & Adolescent Mental Health*, 18 (2), 55 – 60.

Mazzeschi, C. & Lis, A. (1999). The Bender-Gestalt test: Koppitz's Developmental Scoring System administered to two samples of Italian preschool and primary school children. *Perceptual and Motor Skills*, *88*(3/2), 1235-44.

McCallin, R. C. (2006). Test Administration. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 625-652). Mahwah, NJ: Lawrence Erlbaum.

Merino, C., Díaz, M., Zapata, L. & Benites, L. (2006). School psychology in Peru. In S. R. Jimerson, T. O. Oakland & P. T. Farell (Eds.) *The Handbook International of School Psychology*, (pp. 299-307). Oakland: Sage. [USA].

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

* sikayax@yahoo.com.ar

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

Merino, C. (2010b). Investigación preliminar del acuerdo inter-calificadores del Sistema de Calificación Cualitativa para el Test Gestáltico de Bender Modificado. *En revisión*.

Merino, C. & Benites, L. (2011). Evaluación de la confiabilidad en dos grupos de edad, usando el Sistema Cualitativo de Calificación para el Test de Bender Modificado. *Universitas Psicológica, 10*(1), 237-249.

Parush, S., Sharoni, C., Hahn-Markowitz, J. & Katz, N. (2000). Perceptual, motor, and cognitive performance components of Bedouin children in Israel. *Occupational Therapy International, 7*, 216–231.

Pauker, J. D. (1976). A quick-scoring system for the Bender-Gestalt: Interrater reliability and scoring validity. *Journal of Clinical Psychology, 32*(1), 86-89

Pelorosso, A. & Etchevers, M. (2004) Baremos de test Gestáltico Visomotor. *Revista Investigaciones en Psicología, 9*(3), 1-16.

Pett, M. A. (1997). *Nonparametric statistics for health care research*. London: Sage.

Pitman, E. J. G. (1939). A note on normal correlation. *Biometrika, 31*, 9-12.

Rajabi, G. (2009). Normalizing the Bender Visual-Motor Gestalt Test among 6-10 year-old children. *Journal of Applied Sciences, 9*, 1165-1169.

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

Román, M. & Vázquez, C. (1984). *Desarrollo de normas locales de la prueba Bender Gestalt para niños puertorriqueños de kinder a tercer grado*. Tesis de maestría, Facultad de Ciencias Sociales, Universidad de Puerto Rico, Recinto de Ríos Piedras.

Rosenblum, S., Katz, N., Hahn-Markowitz, J., Mazor-Karsenty, T. & Parush, S. (2000). Environmental influences on perceptual and motor skills of children from immigrant Ethiopian families. *Perceptual and Motor Skills, 90*, 587–594.

Shapiro, S. S. & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika, 52*(3/4), 591-611.

Shrout, P. E. & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420-428.

Sisto, F. F., Noronha, A. P. P. & Santos, A. A. A. (2006). *Teste Gestaltico Visomotor de Bender – Sistema de Pontuaçao Gradual (B-SPG)*. Sao Paulo, SP: Vetor.

Smirnov, N. V. (1939). Estimate of deviation between empirical distribution functions in two independent samples. *Bulletin Moscow University, 2*(2), 3-16.

Tolor, A. & Brannigan, G. G. (1980). *Research and clinical applications for the Bender-Gestalt Test*. Springfield, IL: Charles C. Thomas Publishers.

Yousefi, F., Shahim, S., Azvanieh, A., Mehryar, A. H., Hosseini, A. A. & Alborzi, S. (1992). Some normative data on the Bender-Gestalt test performance of Iranian children. *British Journal of Educational Psychology, 62*, 410-416.

* Docente investigador del Instituto de Investigación de la Escuela Profesional de Psicología - Universidad de San Martin de Porres, Perú.

* sikayax@yahoo.com.ar