UN ANÁLISIS NO PARAMÉTRICO DE ÍTEMS DE LA PRUEBA DEL BENDER MODIFICADO PARA ESTUDIANTES DE PRIMARIA

A NONPARAMETRIC ITEM ANALYSIS OF THE BENDER GESTALT TEST MODIFIED FOR PRIMARY STUDENTS

César Merino Soto*
Universidad Científica del Sur

Recibido: 09 de enero de 2009 Aceptado: 05 de mayo de 2009

RESUMEN

La presente investigación hace un estudio psicométrico de un nuevo sistema de calificación de la Prueba Gestáltica del Bender modificada para niños, que es el Sistema de Calificación Cualitativa (Brannigan y Brunner, 2002), en un muestra de 244 niños ingresantes a primer grado de primaria en cuatro colegios públicos, ubicados en Lima. El enfoque usado es un análisis no paramétrico de ítems mediante el programa Testgraf (Ramsay, 1991). Los resultados indican niveles apropiados de consistencia interna, identificándose la unidimensionalidad, y el buen nivel discriminativo de las categorías de calificación de este Sistema Cualitativo. No se hallaron diferencias demográficas respecto al género ni la edad. Se discuten los presentes hallazgos en el contexto del potencial uso del Sistema de Calificación Cualitativa y del análisis no paramétrico de ítems en la investigación psicométrica.

Palabras clave: Prueba gestáltica de bender, sistema cualitativo de calificación, visomotricidad, teoría de respuesta al ítem, testgraf.

ABSTRACT

This research designs a psychometric study of a new scoring system of the Bender Gestalt test modified to children: it is the Qualitative Scoring System (Brannigan & Brunner, 2002), in a sample of 244 first grade children of primary level, in four public schools of Lima. The approach applied is the nonparametric item analysis using the Test graft computer program (Ramsay, 1991). Our findings point to good levels of internal consistency, unidimensionality and good discriminative level of the categories of scoring from the Qualitative Scoring System. There are not demographic differences between gender or age. We discuss our findings within the context of the potential use of the Qualitative Scoring System and of the nonparametric item analysis approach in the psychometric research.

Keywords: Bender Gestalt Test, Qualitative Scoring System, visualmotor, item response theory, Testgraf.

En la investigación educativa y práctica profesional, incluyendo áreas de epidemiología médica, las habilidades de coordinación ojo-mano continúan siendo la variable de respuesta en estudios longitudinales y transversales, por ejemplo al evaluar el impacto del plomo en el desarrollo visomotriz (Azcona, Rothenberg, Schannaas, Romero y Perroni, 2000), o en áreas como la optometría al estudiar correlacionalmente la integración visomotora con el rendimiento académico (Kulp, 1999). Pero el interés no es sólo viene de la investigación básica sino también de la elaboración y diseño de instrumentos. La creación de nuevas herramientas para la evaluación de la visomotricidad que crean un puente entre la precisión y la validez, y la facilidad de aplicación y calificación, está caminando a paso acelerado, tal como se demuestra en los recientes desarrollos de pruebas evolutivamente sensibles y aplicados en espacios profesionales diferentes a la psicología escolar, como en la medicina pediátrica (Pascual, 2001a, 2001b; Bojórquez, 2005)

Aún cuando pueden existir instrumentos de evaluación de la visomotricidad no publicadas, son las publicadas que garantizan un buen soporte psicométrico en su construcción. Herramientas muy conocidas son la Prueba de Integración Visomotora (Beery, 2000) y el Test Gestáltico de Bender (Bender, 1987). El test de Bender es uno de los más populares internacionalmente, y varios sistemas de calificación se han creado. Recientemente creado y revisado, el Sistema de Calificación Cualitativa

* sikayax@yahoo.com.ar
ISSN: 1729 - 4827

(SSC, Brannigan & Brunner, 1989, 1996, 2002) evalúa la exactitud de cada dibujo en una escala de 6 puntos desde 0 hasta 5. Además de las líneas directivas generales, este sistema también provee directivas específicas y ejemplos para acumular puntos cada diseño. Se creó teniendo en mente la evaluación de la calidad global de las reproducciones de niños desde los 4 años, 6 meses hasta los 8 años, 5 meses; esta evaluación es denominada cualitativa o gestáltica. El sistema es similar al recientemente lanzado Sistema Global de Calificación del Bender II (Brannigan & Decker, 2003) y usa el mismo enfoque estricto de calificación que requiere que los dibujos sean tan buenos o mejor que los ejemplos citados en un determinado nivel (Brannigan & Brunner, 2002) para recibir crédito en ese nivel.

El sistema de calificación fue diseñado para usarse con una versión modificada de la prueba original del Bender, que únicamente incluye seis de los más apropiados para predecir logro escolar en niños menores de edad entre 4 años y 8 años (láminas A, 1, 2, 4, 6, 8). Esta modificación provino del trabajo conjunto entre Bender, y Jansky y deHirsh para el índice predictivo de Jansky (Jansky y deHirsh, 1972). Posteriormente, otro sistema como el Sistema Sugar, basado en esta modificación y orientado también al sistema global de calificación, proliferó brevemente (Sugar, 1995; Parsons y Weinberg, 1993) dado quizás a que su aplicación tenía un estrecho rango, es decir, niños que ingresan al primer grado de primaria.

Actualmente el SSC es un nuevo competidor de uno de los sistemas más populares y tradicionales para calificar las reproducciones de las figuras del Bender en niños: el Sistema Evolutivo de Calificación (Koppitz, 1984). El sistema Koppitz ha sido largamente utilizado desde su creación, y ha generado más de 300 estudios publicados (Bollen, 2003) y representa uno de los principales enfoques psicométricos para estimar el funcionamiento visomotor y de ajuste conductual del Test de Bender (Cobrinik, 1988).

Aún hoy continúa enseñándose en las universidades en nivel de pre-grado; y actualmente hay información sobre datos normativos recientes en Argentina (Casullo, 2001) y Estados Unidos (Bolen, 2003) e Italia (Lis y Mazzeschi, 1999; 2000). El sistema de Koppitz consiste en 30 errores discretos que se puntúan cuando ocurren en las reproducciones, asignando 1 si esta presente el error y 0 si no lo está. Desde su publicación original en inglés en 1964, ha sido el sistema de puntuación más preferido, destacándose por que se basa en la evaluación de errores discretos en la reproducción de cada una de las 9 láminas. Sin embargo, la evaluación con este tipo de sistema ha sido criticado dado su sobre simplificación y el examen molecular de los errores en la reproducción de los diseños (Chan, 2000; Brannigan & Brunner, 2002). En tal punto, Lauretta Bender insistía en que el funcionamiento visomotor podría ser capturado más apropiadamente con una evaluación que exigiera examinar globalmente la calidad de la gestalt, y que evitara segmentar esta evaluación (Brannigan y Brunner, 2002), justamente por el estatus de unidad dinámica de su desempeño y que debería ser interpretado integrativamente (Cobrinik, 1988).

Las investigaciones conducidas sobre el sistema Koppitz respecto a los indicadores emocionales y evolutivos son muy frecuentes y sus normas antiguas aún preferidas (Michelle-Burns, 2000), pero el nuevo SSC aún no ha sido beneficiado de tal popularidad. Hasta la fecha, no se ha reportado en el habla hispana análisis de confiabilidad, de validez o normativos del SCC; sólo un estudio en Hong Kong reportó información sobre la confiabilidad, validez y comparaciones normativas (Chan, 2000a, 2000b). Las técnicas de análisis de ítems desde la teoría clásica de los test, por ejemplo, índices de dificultad y discriminación son útiles pero técnicas modernas de análisis como la Teoría de Respuesta al Item (TRI) dan diferentes opciones de análisis, como aquellos obtenidos de los gráficos de función de las respuestas al ítem. Uno de los aspectos que se evalúan en esta teoría es el funcionamiento del ítem, y específicamente de sus opciones de respuesta mediante la curva característica del ítem o de opción (Lei, Dumbar y Kolen, 2004). Estos métodos tienen su espacio interpretativo dentro de modelos paramétricos del TRI, pero aplicar estos métodos debido las sofistificaciones matemáticas, tamaño muestral y formato de los ítems (Sachs et al., 2001). Pero modelos no paramétricos de TRI, que usan técnicas de modelamiento kernel son más flexibles y se ajustan mejor a las condiciones muestras relativamente pequeñas (Ramsay, 1991)

Estimación no paramérica de las curvas características de opción

La estimación no paramétrica de las curvas de opción inicia con el ordenamiento de cada examinado de acuerdo al puntaje obtenido, que luego son convertidos a unidades estandarizadas para estimar el puntaje de atributo latente.

Una serie de ponderaciones ajustando las respuestas de los examinados a una función kernel permite la estimación de cada puntaje en el ítem en una curva estimada de valores del atributo latente (Santor et al, 1994; Ramsay, 1995a). Estas curvas retratan los cambios en la probabilidad de elegir una opción como una función del atributo latente medido. En los ítems de tipo escala, es decir ítems politómicos ordenados, la curva de opción debería elevarse en las opciones de mayor magnitud a medida que aumenta el puntaje de la prueba. De este modo, la curva sugiere que el desempeño de las opciones de respuesta es una función del

* sikayax@yahoo.com.ar

ISSN: 1729 - 4827

atributo medido. En este análisis visual es útil observar el grado de traslape entre las opciones. Si dos curvas se superponen, ello puede sugerir que una mejor precisión de medición se podría obtener si tales opciones de unifican, en lugar de funcionar independientemente. Dado que la descripción del ítem usa su curva de opción característica, se propuso un modelo de teoría de respuesta al ítem no paramétrica y apropiada para moderados tamaños muestrales, basados en el ajuste suavizado kernel (Ramsay, 1991) y conducido por el programa TestGraf (1995a, 2000).

El programa Testgraf provee la presentación de gráficos para examinar cómo funcionan las opciones de respuesta a lo largo del puntaje de la prueba, que representa el atributo medido. En la producción de los gráficos de curvas características de opción, habrá referencias fijas a modo de líneas fragmentadas verticales, que se interpretan como cuantiles sobre el porcentaje de personas que caen en tal posición o debajo de ellas. Adicional a este análisis de las curvas de opción, el programa facilita el examen de la confiabilidad condicional al nivel del atributo, es decir, a lo largo del puntaje de la prueba. Ejemplos representativos del examen de las opciones se han efectuado sobre pruebas relacionadas con el rendimiento metacognitivo (Sachs, Law, Chan y Rao, 2001) y con el Inventario de Depresión de Beck (Santor, Ramsay, Zuroff, 1994).

Función de confiabilidad

La confiabilidad es una estimación del error de medición introducido en los puntajes de una prueba (Nunnally y Bernstein, 1995). De los varios tipo de confiabilidad, la consistencia interna por el coeficiente alfa de Cronbach (Cronbach, 1951) es la aparentemente más reportada. La medida tradicional de calidad de la prueba es este coeficiente de confiabilidad, pero esta es una medida "omnibus" y no muestra cómo la calidad de la prueba varía en función del nivel del atributo medido (Sachs at al, 2001).

Graficar los cambios en la estimación de la confiabilidad clásica, y su expresión individualizada en el error estándar de medición, lleva al usuario a tener más información para evaluar el impacto del error de medición sobre los puntajes en el test del Bender.

La presentación gráfica de la confiabilidad como variable dependiente del nivel de atributo medido tiene una interpretación similar la función de información de un puntaje (Ramsay, 2000), estimada por ajuste suavizado kernel en el programa Testgraf (Ramsay, 1995a, 2000). La función de información del test es el mayor indicador de cómo una medida se desempeña en varios niveles del atributo (Santor & Ramsay, 1998). Dado esto, se considera una medida más útil que el coeficiente alfa de Cronbach, pues nos permite observar cómo varía la precisión de la prueba como una función del atributo latente. No es raro hallar que pocas pruebas se desempeñan bien en todos los niveles del atributo medido. La función de información se interpreta similarmente a la función de confiabilidad.

Regresando al SCC, tal sistema es nuevo en la práctica profesional y no se han reportado estudios que exploren sus características métricas en países del habla hispana; pero hay un emergente interés que está es desarrollado (Merino, en revisión) y cuya variante temática se incluye en este estudio. El presente estudio tiene por objetivo examinar psicométricamente los puntajes a nivel total y a nivel de ítems usando el Sistema de Calificación Cualitativa de Brannigan y Brunner, para la versión modificada del Test Gestáltico de Bender.

En primer lugar, se observará el funcionamiento de las opciones o niveles de respuesta a cada lámina; esto se hará con la curva característica de la opción; los análisis del grado de diferenciación de las opciones es útil ya que revelerá el grado que el Sistema Cualitativo logra separar los niveles de exactitud en las reproducciones usando la escala de 6 puntos. En segundo lugar, se examinará la consistencia interna mediante la función de confiabilidad. Estos análisis se efectuarán con usando el programa Testgraf (Ramsay, 1995a, 2000) que expresa un enfoque no paramétrico del TRI.

También examinaremos demográficamente el impacto de la procedencia educativa de los niños sobre el nivel de puntuación en la prueba pero usando los puntajes esperados y no los puntajes directos; los puntajes esperados se basan en una estimación de máxima verosimilitud del nivel de atributo y es un estimador más exacto del verdadero nivel del examinado sobre el constructo medido (Santor et al., 1994; Sachs et al, 2001)

Método

Participantes

Los participantes de nuestro estudio 244 niños ingresantes al primer grado de educación primaria, distribuídos en 4 colegios públicos situados en la zona urbana de un distrito costero dentro y al sur de Lima. Los colegios se caracterizan por ser unidocentes en el nivel primaria, y contener en cada aula 30 alumnos en promedio. Los datos en la Tabla 1 presentan la información demográfica. La edad promedio de los niños es de 70 meses (de = 5.2), con una mínima edad de 51 hasta 93 meses; las diferencias en la media de edad en cada colegio no ha sido de gran magnitud como separar los análisis. La proporción de varones y mujeres es similar en los colegios participantes y en la muestra total. Teniendo presente la población aparentemente normal desde la cual provienen los niños, únicamente un pequeño porcentaje de

* sikayax@yahoo.com.ar

ISSN: 1729 - 4827

madres reportaron que sus niños recibieron algún tipo de asistencia psicopedagógica en algún momento de la historia preescolar. El nivel modal de estudios de las madres es generalmente de secundaria completa, y aproximadamente menos del 10% tiene estudios superiores completos. Las madres se dedican más frecuentemente a las labores hogareñas y en menor proporción dedicadas a trabajos a tiempo completo o parcial, pero que combinan con actividades independientes para generar ingresos. Por esta misma razón, ls colegios de nuestros participantes tienden a captar familias de nivel socioeconómico que limita con el nivel medio bajo a menos, y de zonas urbanas y urbano-marginales

Usualmente, todos los niños vienen recibiendo un número de años de instrucción preescolar, y excepcionalmente, alguno no ha participado de algún programa preescolar en algún momento. Si la convivencia con ambos padres era formalizada por el matrimonio, casi la tercera parte de los niños conviven con ambos padres y en segundo lugar, únicamente con la madre.

Tabla 1

Descripción demográfica de los participantes

	N	%
Colegio		
C.E.M.I.	96	39.3
C.E.S.M.	93	38.1
C.E.A.R.	13	5.3
C.E.S.J.O.	42	17.2
Sexo		
Varón	141	57.8
Mujer	103	42.2
Asistencia del niño a terapia		
Sí	35	14.3
No	192	78.7
No respondió	17	7.0
Nivel educativo (padres)		
Prim. Incomp.	9	3.7
Prim. Comp.	10	4.1
Sec. Incomp.	40	16.4
Sec. Comp.	85	34.8
Tec. Incomp.	16	6.6
Tec. Comp.	40	16.4
Univ. Incomp.	6	2.5
Univ. Comp.	11	4.5
No describe	27	11.1
Convivencia familiar		
Con ambos padres	159	65.2
Solo la madre	52	21.3
Solo el padre	8	3.3
Con otros	1	.4
No describe	24	9.8
Mes de evaluación		
1er.	40	16.4
2do.	47	19.3
3ro.	82	33.6
4to.	75	30.7
Total	244	244

Instrumento

Test Gestáltico de Bender Modificado. La versión modificada 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 son los más apropiados para niños pequeños. El manual describe un sistema para puntuar el desempeño gráfico del niño, el Sistema de Calificación Cualitativa, 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 inter-jueces, y la validez del Sistema Cualitativo de Calificación da soporte a sus propiedades métricas y sus cualidades intrumentales en la evaluación psicopedagógica (Brannigan & Brunner, 2002). Frente al Sistema Evolutivo de Calificación de Koppitz, el SCC muestra correlaciones más elevadas con criterios de rendimiento escolar en el estudio original (Brannigan & Brunner, 2002) 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 inter-examinadores son satisfactorios. En nuestro estudio, el coeficiente de acuerdo intraclase entre tres examinadores usando una muestra aleatoria de 25 protocolos fue 0.71, que es considerado de buen nivel de acuerdo (Merino, 2006)

Procedimiento

La recolección de datos se efectuó en el contexto de la convocatoria recibir matrícula de niños para el ingreso a

* sikayax@yahoo.com.ar

ISSN: 1729 - 4827

primer grado. Al momento de la evaluación, no se detectó niños con discapacidades cognitivas severas, así como otras discapacidades que hubieran podido ser detectadas por conductas atípicas durante el rendimiento. Todos los niños fueron acompañados por sus madres y o apoderados; mientras se evaluaban a los niños en un aula, paralelamente los padres llenaron cuestionarios en otra aula; los padres llenaron un cuestionario demográfico. Los niños fueron evaluados con una batería de pruebas que incluía el dibujo de la figura humana (Reynolds y Hickman, 2004), una prueba de despistaje de habilidades para primer grado (Merino, 2007) y la versión grupal del test de Bender. Esta versión grupal requirió de cuadernillos en que cada figura estuvo impresa en cada página, exactamente en el tercio superior de la hoja. Dos examinadores en cada aplicación explicaron en qué consistía la tarea y se mantuvieron las recomendaciones estándares sugeridas por el manual.

Tabla 2

Estadísticos básicos para los ítems y correlaciones inter-ítem de la prueba de Bender

Láminas	M	DS	A	1	2	4	6	8
A	3.26	.88	1					
1	3.30	1.03	.381**	1				
2	3.04	.85	.341**	.619**	1			
4	3.39	.77	.445**	.373**	.328**	1		
6	3.26	.86	.415**	.454**	.393**	.617**	1	
8	3.23	.91	.428**	.475**	.418**	.444**	.555**	1

** P < 0.01 (bilateral)

Tabla 3

Estadísticos descriptivos básicos y confiabilidad alfa de Cronbach (α) y sus intervalos de confianza (95%)

	Media	D.E.	α (I.C. 95%)
Colegio			
C.E.M.I.	20.06	4.036	0.80 [0.73, 0.85]
C.E.S.M.	18.88	3.590	0.78 [0.70, 0.84]
C.E.A.R.	17.08	5.634	0.94 [0.87, 0.97]
C.E.S.J.O.	20.02	2.884	0.79 [0.67, 0.87]
Sexo			
Varón	19.8	3.6	0.78 [0.71, 0.83]
Mujer	18.9	4.4	0.85 [0.80, 0.89]
Total	19.45	3.85	0.81 [0.77, 0.84]

Resultados

Diferencias demográficas. Usando las estimaciones de atributo latente, no se detectaron diferencias estadísticamente significativas (t [242] = 1.86, p = 0.06) en el rendimiento visomotor entre niños (M = 19.8, DE = 3.63) y niñas (M = 18.9, DE = 4.09). Luego, la diferencia en el desempeño visomotriz entre los niños que asistieron a algún tipo de terapia frente al resto tampoco fue mayor de lo que se puede haber producido por error de muestreo, t(225) = 0.95, p = 0.34.

Comparando los colegios desde el cual provinieron los alumnos, el ANOVA una vía rechazó la hipótesis nula de igualdad de medias, F(3, 240) = 3.54, p = 0.01; una comparación post hoc ajustando el nivel de significancia por el método Bonferroni detectó diferencias marginalmente significativas (p = 0.05) provenientes únicamente del colegio A.R. (M = 17.08, DE = 5.63) frente al colegio M.I (M = 20.06, DE = 4.03), pero con una magnitud moderadamente alta (d Cohen = 0.71). Esto nos sugiere que es posible detectar rendimientos diferentes entre-grupos, pero que intragrupalmente son homogéneamente bajos en el funcionamiento visomotor.

Por otro lado, la correlación lineal entre la prueba Bender y la edad de los niños fue -0.08 (p > 0.05), que nos indica que los efectos de la edad sobre el desempeño visomotor provienen por variaciones del muestreo y no por diferencias sistemáticas respecto a la edad en el rango evaluado.

Evaluación psicométrica

Unidimensionalidad. Como en un reporte preliminar anterior (Merino y DeRoma, en prensa), la varianza

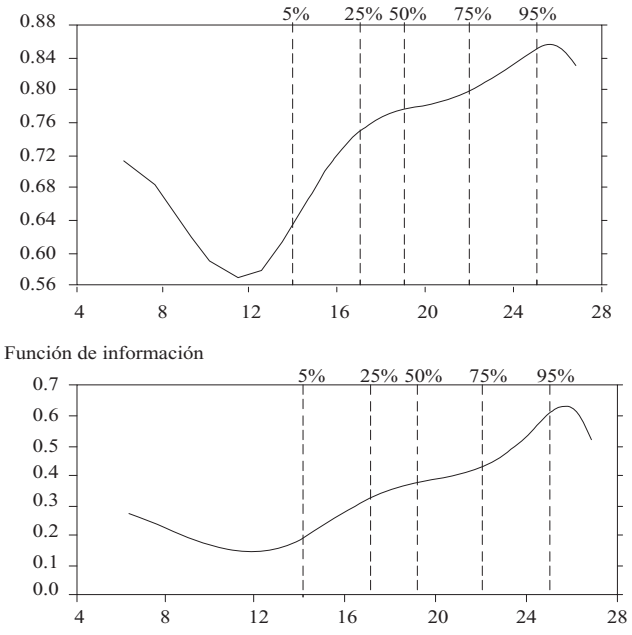

Figura 1: Parámetros de estimación del error de medición

*sikayax@yahoo.com.ar

ISSN: 1729 - 4827

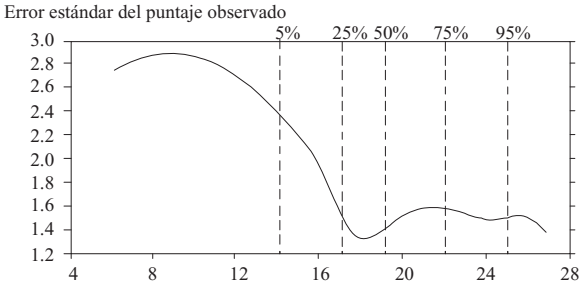

Figura 1
Parámetros de estimación del error de medición

explicada por el único componente (53.9%) es cuantitativamente similar lo hallado anteriormente, 47%.

Esta estimación de la dimensionalidad, obtenida por la extracción de un autovalor de la matriz de correlaciones inter-ítem, sugiere que un solo factor parsimoniosamente está presente en la definición latente del constructo de integración visomotora. La integridad del constructo representado se mantiene por lo tanto, constante en este estudio que ha utilizado participantes diferentes.

Consistencia interna. El coeficiente alfa de Cronbach para los puntajes se ha mantenido dentro niveles moderadamente altos. Para la muestra total, la consistencia interna está alrededor de 0.82, y tal es una magnitud de buen nivel dentro del esquema indicado por Ciccheti (1994). Similares valores se han hallado en Brannigan y Brunner (2002). Entre los colegios, se ha observado variabilidad en el grado de error de medición (desde 0.77 hasta 0.87), pero estas variaciones no han sido lo suficientemente grandes como para declarar una diferencia sistemática y significativamente estadística entre alguna de ellas.

Al comparar los valores de la confiabilidad entre varones y mujeres, los primeros tienden a dar respuestas más confiables que las niñas (0.85 vs. 0.77); no nos es claro la razón de estas diferencias en al confiabilidad. La homogeneidad de los ítems ha sido óptima, ya que la correlación inter-ítem promedio 0.44 y desde 0.03 hasta 0.61 para la muestra total; este nivel está dentro del rango que refleja medidas que evalúan constructos de amplio espectro (Clark y Watson, 1995). De manera similar, las correlaciones ítem-test están en un nivel promedio y rango bastante aceptables. En la Tabla 2 se presentan estos valores, además de los obtenidos de acuerdo al colegio y al sexo.

Teniendo en cuenta el nivel de las reproducciones de los niños y la calificación de los examinadores, los valores promedio para las seis figuras se hallan alrededor del punto 3 (Tabla 2); y la variabilidad de las calificaciones ha sido mayor en la lámina 1 (d.e. = 1.03); en el resto, la variabilidad ha demostrado valores cercanos entre sí.

Sin embargo, las confiabilidades estimadas mediante el coeficiente alfa de Cronbach (Cronbach, 1951) y su estimación del error individual, el error estándar de medición, son medidas globales o estáticas (Sachs, et al., 2001).

Desde la Teoría de Respuesta al Ítem, la función de información ofrece una mejor observación de la confiabilidad a lo largo de los niveles de habilidad definidos por el puntaje obtenido en el sujeto; pero una curva de la confiabilidad a lo largo de los niveles de rendimiento visomotor es equivalente y más familiar para el lector. En nuestro estudio, la curva de la función de confiabilidad muestra un patrón irregular de precisión a lo largo de los puntajes. Su más baja estimación (0.57) se halla en cerca del percentil 5 en la muestra (puntaje directo = 11.5), y rápidamente aumenta hasta el primer cuartil.

Luego se estabiliza para seguir aumentando lentamente hasta su pico cerca del percentil 95 (0.84); se observa que después del percentil 75% se puede lograr una confiabilidad mínima de 0.50. El recorrido de la función de información describe un patrón visualmente similar pero suavizado en su incremento monotónico: más información relevante al constructo se obtendrá en niveles elevados del desempeño motriz. Ambos gráficos concuerdan que la precisión de la medición varía en un amplio rango que va desde lo inaceptablemente bajo hasta uno moderadamente alto.

Menos precisión se obtiene en los niveles bajos del atributo medido y, por el contrario, mejor precisión se consigue después del primer cuartil. El error estándar de medición, sin embargo, alcanza su pico cerca de la puntuación promedio y disminuye ligeramente después de este centro; esto puede ser efecto de la menor dispersión de los puntajes observada encima de una desviación estándar de la media. La puntuación individual, por lo tanto, será más variable e imprecisa en tales niveles de puntuación.

Curvas características de opción. La progresión de las opciones en cada nivel del atributo ha sido bien diferenciada, ya que los diferentes cuantiles se ha observado que las curvas de opción extremas han seguido un patrón esperable en tales niveles de atributo (ver Figuras 2, 3, 4 ,5 ,6 ,7). Por ejemplo, las opciones 0 y 1 generalmente se han mantenido debajo del primer quintil, mientras que los niveles de puntaje 4 y 5 han tenido su pico en el cuarto y quinto quintil respectivamente. Los puntajes superiores más extremos prácticamente han provenido del puntaje 5, mientras que la frecuencia del puntaje 4 decrecía en este nivel de atributo. La lámina A no recibió algún puntaje de 2, y ello puede sugerir que esta las reproducciones o los examinadores no capturan apropiadamente este nivel de desempeño (Figura 2).

La observación de las curvas de opción también nos lleva a detallar que cada nivel de puntaje parece ser dominante en los cuantiles, y que tal dominancia crece o

* sikayax@yahoo.com.ar

ISSN: 1729 - 4827

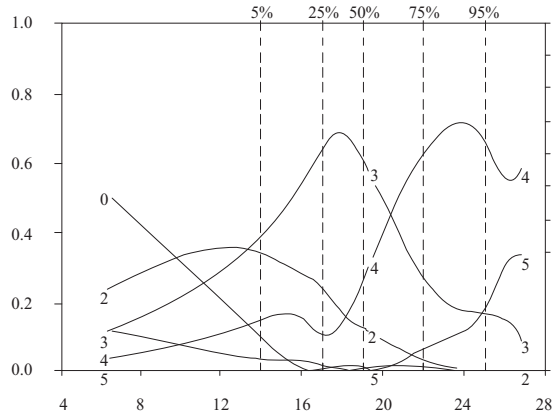

Figura 2
Curvas de probabilidad de las opciones de respuesta de la lámina A

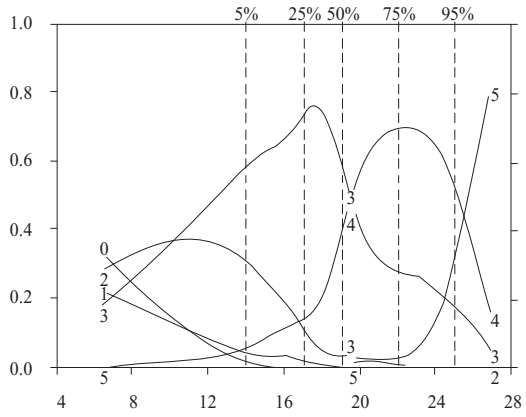

Figura 5
Curvas de probabilidad de las opciones de respuesta de la lámina 4

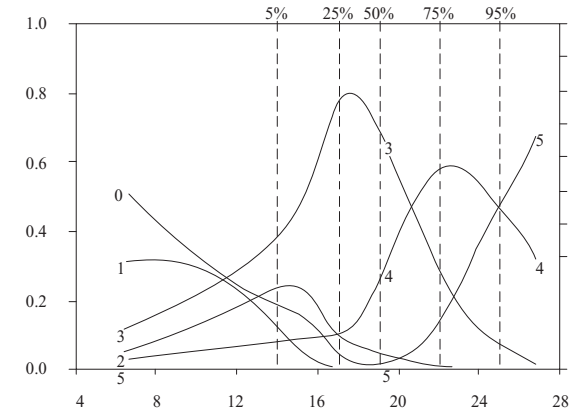

Figura 3
Curvas de probabilidad de las opciones de respuesta de la lámina 1

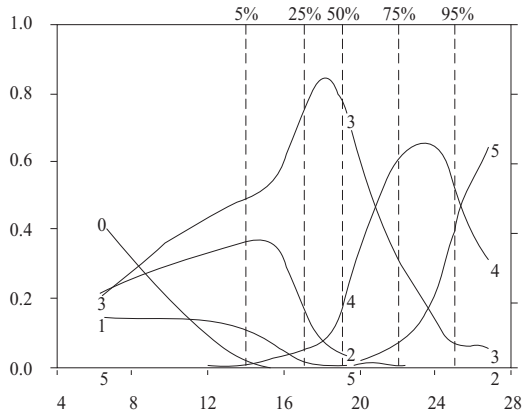

Figura 6
Curvas de probabilidad de las opciones de respuesta de la lámina 6

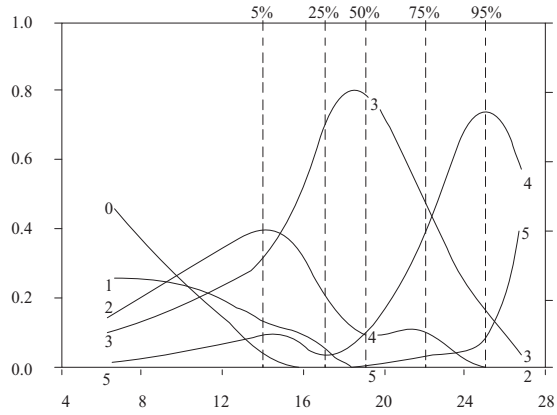

Figura 4
Curvas de probabilidad de las opciones de respuesta de la lámina 2

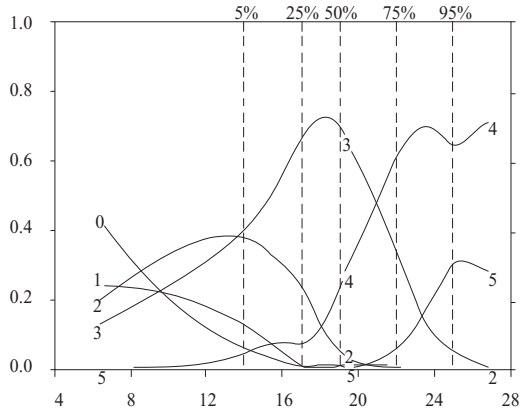

Figura 7
Curvas de probabilidad de las opciones de respuesta de la lámina 8

*sikayax@yahoo.com.ar

ISSN: 1729 - 4827

decrece según avanza en el nivel de atributo. La diferenciación de estos niveles de puntaje en cada lámina sugiere que existe un ordenamiento ideal de estos niveles en el rango total de puntaje del test de Bender con el Sistema de Calificación Cualitativa; debido a esta diferenciación, se puede asumir que los puntajes discriminan apropiadamente rendimientos desde un nivel bajo hasta uno de elevado rendimiento.

Discusión

En el presente estudio nuestro objetivo ha sido examinar las propiedades de confiabilidad y en funcionamiento de los ítems de la versión modificada del Test de Bender usando el Sistema de Calificación Cualitativa; el método estadístico es un enfoque relativamente nuevo, basado en el análisis funcional de datos y en la teoría no paramétrica de respuesta al ítem (Santor et al, 1994; Ramsay, 1995a).

La estimación de las propiedades métricas de los ítems consistió en hallar la mejor descripción de su funcionamiento a lo largo de los variados niveles de habilidad de un sujeto o grupo de sujetos; que es una situación de ajuste a una curva típico del llamado análisis funcional de datos (Rossi, Wang, y Ramsay, 2002); la curva de interés que describe el ajuste o el modelamiento de los datos es el funcionamiento de la respuesta del ítem.

El impacto de este enfoque desde sus características puede ser importante para su inclusión en las estrategias de análisis de ítems, ya que la medición basada en el modelamiento de datos, como la teoría de respuesta al ítem, está popularizándose y es el objetivo de la teoría moderna de los tests (Ramsay, 1991). El enfoque específico utilizado fue de tipo no paramétrico (Ramsay, 1991), que es metodológica y computacionalmente atractiva por su flexibilidad, tal como ocurre en las aplicaciones no paramétricas inferenciales que típicamente se enseñan en los cursos para estudiantes no graduados. Al extender su uso en el estudio del sesgo de los ítems, se ha hallado que su poder de detección del funcionamiento diferencial de ítems han sido eficiente para su uso en muestras de pequeño a moderada tamaño (Zumbo, & Witarsa, 2004), y que los gráficos producidos por la función no paramétrica kernel son excelentes puntos de análisis para determinar diferentes formas de funcionamiento diferencial de ítems (Xuan y Gierl, 2005).

Viendo los resultados respecto a la consistencia interna y la función de confiabilidad en el test de Bender, las magnitudes de la confiabilidad estimada mediante el coeficiente alfa (Cronbach, 1951) han sido generalmente apropiadas para esta medida caracterizada por ser un instrumento de despistaje de las habilidades visomotoras y considerando también el reducido número de ítems. Las diferencias entre la consistencia interna de las submuestras por colegio y género no han sido en general grandes, excepto para el colegio A.R. Hallar a un grupo con problemas en la interpretación de sus puntajes basados en la baja consistencia interna debe advertir al investigador y al profesional sobre valorar este dato psicométrico en su práctica.

Además, las interpretaciones que haga deben ser moderadas por los niveles de error de medición variables en los grupos de participantes que como en nuestro estudio hemos hallado. La imprecisión de estas estimaciones de confiabilidad se ha reflejado en el pequeño tamaño entre las submuestras, ya que la amplitud del intervalo del 95% de confianza ha recorrido desde los niveles altos a moderadamente bajos de consistencia interna. Se requiere un tamaño muestral mayor para hacer una estimación más precisa de la consistencia interna, y las recomendaciones actuales sugieren 400 como un tamaño apropiado (Charter, 1999)

Las variaciones de la consistencia interna, revelada a través de los gráficos de la función de confiabilidad sugieren que este aspecto de la calidad de los instrumentos no es estático y sí vulnerable a los aspectos idiosincrásicos de los grupos muestrales en análisis, y esto está de acuerdo con las recomendaciones y estándares modernos para explorar la confiabilidad (AERA, APA y NCME, 1999; Onwuegbuzie y Daniel, 2002) que recomiendan estimar la confiabilidad no únicamente para la muestra total sino para los subgrupos que las componen. Aunque las diferencias de confiabilidad halladas no han sido sugestivas de problemas en la homogeneidad de las correlaciones entre los ítems, otro aspecto parece ser problemático para la interpretación de nuestros resultados.

La amplitud de los intervalos de confianza permitió traslapes entre los intervalos calculados, pero en condiciones de mayor tamaño muestral, las estimaciones de confiabilidad obtenidas hubieran sido detectadas como estadísticamente significativas, ya que estas intervalos se estrecharían. Pero la variabilidad de la consistencia interna también ha provenido del nivel de puntaje o atributo medido.

De este modo, del análisis de la función de confiabilidad observamos también que más información relevante al constructo se obtendrá en niveles elevados del desempeño visomotor, y una mayor presencia del error ocurre en los niveles bajos del atributo; este es una situación que requiere ser confirmada en otro grupo de participantes, ya que el impacto en el uso de la prueba es importante. Si un instrumento es menos confiable en el nivel bajo del atributo medido, el profesional debería elegir otro instrumento que le permita obtener resultados más precisos en la detección y diagnóstico de problemas visomotores. Podemos concluir

* sikayax@yahoo.com.ar

ISSN: 1729 - 4827

que la consistencia interna para nuestros resultados tiende a ser apropiada para fines de evaluación de grupo, y ha resultado ser moderadamente variable entre los distintos grupos de la muestra

Si esta imprecisión proviene de la inconsistencia de los calificadores del Sistema Cualitativo, una mejor preparación del uso de los criterios de calificación mejorará la precisión de la evaluación de las reproducciones que reflejen déficits en la integración visomotora. Sesiones de entrenamiento entre los calificadores deben poner más atención a los protocolos de niños con desempeños notoriamente bajos. Paralelamente, el bajo desempeño visomotor está relacionado con el bajo rendimiento escolar (Köppitz, 1984; Kulp, 1999; Beery, 2000; Brannigan y Brunner, 2002) y la presición del diagnóstico ayudado por la versión modificada del Test de Bender debe ser nuevamente evaluada.

En la detección de problemas de aprendizaje no-verbales, se puede advertir un bajo rendimiento en el Bender, considerando que un signo típico es la discrepancia confiable del rendimiento de la cognición espacial frente a otras medidas cognitivas verbales (Pennington, 1991). Pero la baja confiabilidad puede provenir de un rendimiento inconsistente en los puntajes ubicados en el primer cuartil, así como una estrategia descuidada, impulsividad u otros aspectos que contaminan la medición de la integración visomotora en estos niveles de atributo.

Al explorar los ítems mediante el funcionamiento característico de sus opciones, hemos hallado que estos siguen un patrón que favorece el poder discriminativo del cada ítem. Las opciones de respuesta han sido independientes y diferenciadas por los examinadores, y por lo tanto, el Sistema Cualitativo de Calificación permite discriminaciones de la calidad de las reproducciones de los niños en cada uno de los diseños.

Los ítems y sus opciones han tendido a funcionar bien, aunque algunos ítems (2 y 6) han tendido a ser menos proclives a recibir puntuaciones elevadas; esto puede provenir de la dificultad inherente de estos diseños o de la estrictez de los calificadores. La evaluación del funcionamiento de los niveles de calificación en cada ítem es, sin embargo, favorable, y permiten diferenciar evolutivamente la calidad de las reproducciones. Todas las opciones de puntaje tuvieron curvas características asociadas a los cambios monotónicos del atributo medido.

Hemos visto que el uso y la comparación de las curvas de opción característica producidas por el enfoque no paramétrico (específicamente desde el programa Testgraf, Ramsay, [2000]) ofrece una perspectiva menos estática del funcionamiento métrico de los ítems, considerando que esta técnica da buenos resultados en condiciones de pequeña muestra frente a los métodos paramétricos más comunes (Lee, Chen y Gugga, 2005). Aunque los resultados gráficos no pueden ser evaluados estadísticamente, como los revisados las curvas de opción característica y la función de confiabilidad, proveen un punto de inicio para posteriores análisis basados en las características de los gráficos, que describen la función característica de los ítems y la confiabilidad; esta información no se obtendría con el cálculo estático de la correlación ítem-test (discriminación del ítem) o la confiabilidad alfa de Cronbach.

La elegancia de este análisis no paramétrico proviene de que la unidad de análisis pasa a ser los ítems y su funcionamiento más que el puntaje obtenido de la suma de los ítems, descrito por medio de la relación no lineal y probabilística del ítem con la variable latente; como método, prueba ser superior a los métodos tradicionales (Ramsay, 1995b), y supera los problemas se usar métodos paramétricos que requieren el cumplimiento estricto de presupuestos y la obtención de grandes muestras (Ramsay, 1995b; Sachs et al, 2001).

Este método computacionalmente complejo es resuelto por el uso de programas como Testgraf (Ramsay, 2000), y proporciona una herramienta de progresiva aceptación y difusión, además de recomendado uso como herramienta interpretativa-diagnóstica de los ítems (Lei, at al., 2004). Revisiones de introducción a este método en áreas diferentes a la psicología y medición educativa ya se están conociendo, por citar unos ejemplos, en administración (Laroche, 2004), aplicaciones en medicina sexual (Sills et al., 2005) o en la metodología de evaluación de segundo idioma (Brisay, 1992).

Tenemos que resaltar una pregunta: ¿es posible diferenciar grupos homogéneamente bajos de habilidad? La respuesta desde nuestros resultados es afirmativa, ya que la variabilidad no ha ocurrido en la consistencia interna sino también en los niveles de puntaje. En uno de los colegios evaluados, el desempeño visomotor ha sido inferior al resto, con una diferencia estandarizada moderadamente baja; esta sola evidencia es suficiente para iniciar inmediatos planes de intervención aprovechando los recursos disponibles.

En nuestro estudio, los niños de bajo rendimiento provinieron todos de un mismo colegio, y aparentemente matriculados por un proceso de auto-selección de las familias con niños expresando problemas en el funcionamiento social y académico. Una exploración sensible a este hecho debe ser propuesto junto con la evaluación de habilidades para el rendimiento escolar y ajuste social en un grupo similar.

Debido que las correlaciones predictivas del funcionamiento visomotor con el rendimiento escolar ha sido consistentemente revelados (Köppitz, 1984; Kulp, 1999; Beery, 2000; Brannigan y Brunner, 2002), el uso de este sistema cualitativo de calificación para el Bender

* sikayax@yahoo.com.ar

modificado será potencialmente útil en los programas de detección temprana de problemas del fracaso escolar. La triangulación con medidas que capturen información desde el padre y del profesor definitivamente mejorará el poder predictivo de la detección temprana.

Finalmente, debemos precisar que la tecnología actual en la evaluación psicológica parece apuntar hacia el desarrollo de sistemas de calificación global, como el del presente estudio, ya que presentan mayores posibilidades de correlaciones elevadas para predecir el desempeño con criterios de rendimiento académico (Brannigan, Decker, & Madsen, 2004), funcionamiento cognitivo (Brannigan & Decker, 2003), funciones de personalidad (Lilienfield, Word y Garb, 2001) u observación conductual (Glutting y Oakland, 1993).

Recientemente, Simmer también propuso un sistema cualitativo basado en 3 puntos para el uso de 8 diseños para tareas de copiado, que mejor predicen el rendimiento escolar en primer grado (Simner, 1994). Por lo tanto, esta estrategia de evaluación es un fuerte competidor contra los sistemas más moleculares, y potencialmente más útil para la creación de instrumentos más sensibles de la conducta en áreas de interés para el investigador y usuario profesional.

Pensamos que la inclusión de un nuevo sistema evaluativo de la visomotricidad como el analizado aquí debería reemplazar los enfoque antiguos que conducen también a usar normas antiguas que cuestionada aplicabilidad.

Referencias

AERA, APA & NCME (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Bender, L. (1987) *El test guestáltico visomotor*. Buenos Aires: Paidós.

Bojórquez, M. (2005) Validación de test grafomotor en población escolar normal de Lima. *Anales de la Facultad de Medicina Lima*, 66(3), 218-224.

Bollen, L. M. (2003) Constructing local age norms based on ability for the Bender-Gestalt Test. *Perceptual and motor skills, 97*(2), 467-476.

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 Publishing.

Brannigan, G. G., & Brunner, N. A. (1996). *The Modified Version of the Bender-Gestalt Test for Preschool and Primary School ChildrenRevised.* 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.* Springfield, IL: Thomas.

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

Brannigan, G. G., Decker, S. L., & Madsen, D. H. (2004). *Innovative features of the Bender-Gestalt II and expanded guidelines for the use of the Global Scoring System.* (Bender Visual-Motor Gestalt Test, Second Edition Assessment Service Bulletin No.1). Itasca, IL: Riverside Publishing.

Brisay, M. D. (**March, 1992**) *Applications of TESTGRAF in Setting Cut-off Points on ESL* Tests. Fourteenth Annual Language Testing Research Colloquium, **Vancouver, British Columbia.**

Casullo, M. M. (1991) *Test de Bender: Normas regionales*. Buenos Aires: Guadalupe

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.

Chan, 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. (1999) Sample size requirements for precise estimates of reliability, generalizability, and validity coefficients. *Journal of Clinical and Experimental Neuropsychology, 21*, 559-566.

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

Clark, L. A. & Watson, D. (1995) Constructing validity:

* sikayax@yahoo.com.ar

Basic issues in objective scale development. *Psychological Assessment, 7*(3), 309-319.

Cobrinik, L. (1988) The Bender Gestalt Test in childhood emotional disorder. *Psychiatric Quaterly, 59*(3), 235-243.

Glutting, J., & Oakland, T. (1993). *GATSB: Guide to the assessment of test session behavior.* Orlando: The Psychological Corporation.

Jansky, J. J. & deHirsh, K. (1972) *Preventing Reading Failure: Prediction, Diagnosis, Intervention.* NY: Harper & Row.

Köppitz, E. M. (1984). *El test gestáltico visomotor para niños* (10ma. ed.). Bs. As: Guadalupe.

Laroche, M. (2004) Analyses traditionnelles et FDI des échelles de mesure : application à l'échelle de l'intensité du raisonnement cognitive. *Canadian Journal of Administrative Sciences, 12*, 259-266.

Lee, Y.-S., Chen, T., & Gugga, S.S. (2005). *A comparison between TestGraf and MULTILOG in the estimation of item parameters and ICC estimates.* Paper presented at the annual conference of the American Educational Research Association, Montréal, Canada

Lei, P., Dumbar, S. B. & Kolen M. J. (2004) A comparison of parametric and non-parametric approaches to item analysis for multiple-choice tests. *Educational and Psychological Measurement, 64*(3), 1 23.

Lilienfield, S. O., Wood, J. M, Garb, H. N. (2001). What's wrong with this picture? *Scientific American, 284*(5), 80-87.

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

Lis A., & Mazzeschi, C. (1999). The Bender Gestalt Test: Koppitz's developmental scoring system administered to two samples of Italian preschool and primary school children. *Perceptual & Motor Skills, 88*, 1235-1244.

Merino, C. (2007) *Batería de Despistaje para Primer Grado.* Instrumento no publicado. Lima: Autor.

Merino, C. (en revisión) El Sistema de Calificación Cualitativa para la Prueba Gestáltica de Bender Modificada: Estudio preliminar de sus propiedades psicométricas. *Personas.*

Merino, C. (2006, Octubre) *Confiabilidad inter-jueces del Sistema de Calificación Cualitativa del Test Gestáltico de Bender para Niños.* Ponencia presentada en el II Congreso Iberoamericano de Psicología, Universidad Gracilazo de la Vega, Lima.

Mitchelle -Burns, J. (2000). Performance in children with and without learning disabilities on Canter's Background Interference Procedure and Koppitz' Scoring System for the Bender test. *Perceptual and Motor Skills, 90*, 875-882.

Onwuegbuzie, AJ, & Daniel, LG (2002). Uses and misuses of the correlation coefficient. *Research in the Schools, 9*(1), 73-90.

Parsons, L. & Weinberg, S. L. (1993). The Sugar Scoring System for The Bender Gestalt Test: An Objective Approach that Reflects Clinical Judgment. *Perceptual and Motor Skills, 77*, 883-893

Pascual, S. I. (2001a) Evaluation of maturity in drawing in childhood. I: Evaluation and validation of a graphomotor test in a population of normal children. *Revista Neurología, 33*(9), 812-25.

Pascual, S. I. (2001b) Evaluation of maturity in drawing in childhood. II: Development and validation of a graphomotor test in a child with neuropsychiatric disability. *Revista de Neurología, 33*(10), 938-47.

Pennington, B. F. (1999) Diagnosing Learning Disorders: A neuropsychological framework. New York: Guilford.

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika, 56*(4), 611-630.

Ramsay, J.O. (1995a). *TESTGRAF: A program for the graphical analysis of multiple choice test and questionnaire data.* Montreal: McGiIl University.

Ramsay, J.O. (1995b). *Some notes on the statistical analysis*

* sikayax@yahoo.com.ar

of tests. Montreal: McGiIl University.

Ramsay, J. O. (2000). *TESTGRAF: A program for the graphical analysis of multiple choice test and questionnaire data.* Department of Psychology. McGill University.

Reynolds, C. R., & Hickman, J. A. (2004). *Draw-A-Person Intellectual Ability Test for Children, Adolescents, and Adults (DAP: IQ).* Austin, TX: Pro-Ed.

Rossi, N., Wang, X. and Ramsay, J.O. (2002) Nonparametric item response function estimates with the EM algorithm. *Journal of the Behavioral and Educational, 27*(3), 291-317.

Sachs, J., Law, Y. K., Chan, C. K., & Rao, N. (2001). A nonparametric item analysis of the Motivated Strategies for Learning Questionnaire-Chinese Version. *Psychologia - An International Journal of Psychology in the Orient, 44*(3), 197-208.

Santor, D.A., Ramsay, J.O., & Zuroff, D.C. (1994). Nonparametric item analyses of the Beck depression inventory: Evaluating gender item bias and response option weights. *Psychological Assessment, 6*, 255-270.

Sills, T., Wunderlich, G., Pyke, R., Segraves, R.T., Leiblum, S., Clayton, A., Cotton, D., and Evans, K. (2005) The Sexual Interest and Desire InventoryFemale (SIDI-F): Item response analyses of data from women diagnosed with hypoactive sexual desire disorder. *Journal of Sex Medicine, 2,* 801818.

Simner, M.L. (1994) Improving the predictive validity of geometric-design copying tasks on instruments used to evaluate school readiness. In C. Faure, P. Keuss, G. Lorette, and A. Vinter (Eds), *Advances in Handwriting and Drawing: A Multidisciplinary Approach.* (pp. 489-499). Paris: Europia.

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

Xuan, T. & Gierl, M. J. (2005, Abril). *Using Global and Local DIF Analyses to Assess DIF across Language Groups.* Paper presented at the annual conference of the NCME, Montreal, Quebec, Canada.

Zumbo, B. D., Witarsa, P. M. (2004). *Nonparametric IRT methodology for detecting DIF in moderate-to-small Scale Measurement: Operating Characteristics and a comparison with the Mantel Haenszel.* Edgeworth Series in Quantitative Behavioural Science. 2004 AERA / NCME.

* sikayax@yahoo.com.ar