

Dibujo de la Figura Humana: Análisis del Funcionamiento Diferencial de los Criterios

Fermino Fernandes Sisto¹

Universidade São Francisco, Itatiba, Brasil

Compendio

Esta investigación ha tenido como objetivo la determinación del funcionamiento diferencial del ítem (DIF) en el test del Dibujo de la Figura Humana utilizado para la estimativa de la inteligencia en niños, considerando la variable sexo. Han sido estudiados protocolos de 2508 estudiantes de la enseñanza primaria y preescolar, cuyo promedio de edad fue de 8,14 años. Gran parte de ellos frecuentaban escuelas públicas (72,1 %) y los grupos para el estudio de DIF fueron compuestos por 1248 niños y 1260 niñas. Los parámetros métricos de los ítems fueron determinados según el modelo logístico de un parámetro de Rasch. Los resultados han indicado la presencia de DIF en 20 de los 46 ítems analizados, siendo que nueve de ellos han sido de fácil ejecución para las niñas y 11 para los niños.

Palabras clave: funcionamiento diferencial del ítem (DIF); dibujo de la figura humana; evaluación psicológica; Teoría de Respuesta al Ítem (TRI).

Human Figure Drawing: An Analysis of the Differential Functioning of the Criteria

Abstract

This research aimed the determination of the differential item functioning (DIF) in the Human Figure Drawing used for the intelligence assessment in children, taking into account the sex variable. A sample of 2508 kindergarten and elementary school students, whose average age was 8.14 years, was studied. The majority of them was composed by students from public schools (72.1%) and the groups for the study of DIF were composed by 1248 men and 1260 women. The metric parameters of the items were determined according to the Rasch model of one logistic parameter. The results indicated the presence of DIF in 29 of 46 analyzed items; more specifically nine items were easy for the girls and 11 for the boys.

Keywords: Differential items functioning (DIF); human figure drawing; psychological assessment; item response theory (IRT).

La primera revisión del sistema de Goodenough para la evaluación del dibujo de la figura de un hombre fue propuesta por Harris en 1963 (Harris, 1991) y es conocido como el test del Dibujo de la Figura Humana de Goodenough-Harris. Ha ampliado el sistema de puntaje de 51 ítems para 73 interpretándolo como una medida de la madurez conceptual. Se ha solicitado que las personas dibujasen una mujer y a si mismas. Asimismo, la tendencia de las niñas para ejecutar consistentemente mejor que los niños se mantuvo en sus datos. El manual relata una gran cantidad de estudios de correlación y los coeficientes han variado entre 0,27-0,72. En general presentaron concordancia con los estudios relatados por Goodenough (1926).

Por su parte, el sistema Koppitz (1968) solicita que el niño dibuje apenas una persona y su sistema de corrección incluye 30 ítems. Ha conservado integralmente 15 de los ítems de Goodenough (1926) y los otros resultaron de cambios hechos en los originales. En lo que atañe a la evidencia de validez, una buena parte de ella es dada por inspección visual de los resultados y se refiere al aumento de la frecuencia de ocurrencia en función de la edad, efectos

del aprendizaje y de la madurez, y en relación a los puntajes de los tests WISC y Stanford-Binet.

Naglieri (1988) también se propuso a desarrollar un sistema de puntaje con normas actualizadas para evaluar dibujos de niños y adolescentes. Ha propuesto cuatro categorías y criterios para evaluar los dibujos hechos independientemente de que fueran un hombre, una mujer y de si mismo. En cuanto a la validez de constructo, el autor ha presentado dos estudios en los cuales ha correlacionado sus puntajes con el sistema Goodenough-Harris, cuyos coeficientes han variado entre 0,75 y 0,87. Ha hallado incluso correlaciones con instrumentos de capacidad académica e intelectual, que fueron significativos, a pesar de muy bajos. Al estudiar las diferencias entre sexos ha quedado evidente la diferencia en las tres figuras dibujadas.

La literatura sobre el dibujo de la figura humana es bastante aventajada. Hay muchos estudios que investigaron su uso y en los Estados Unidos ya estuvo entre los tres más usados (Louttit & Browne, 1947), o en octavo (Brown & McGuire, 1976). Evaluando la utilización de tests por área Goh, Teslow y Fuller (1981) han hallado que el test de Goodenough fue el segundo más citado por el criterio de uso.

La detallada revisión (de 1963 la 1977) de Scott (1981) de la propuesta de Goodenough-Harris (Harris, 1991) ha

¹ Dirección: Rua Carlos Guimarães, 150 ap 82, 13024-200, Cambui, Campinas, São Paulo, Brasil. E-mail: fermino.sisto@gmail.com

enseñado algunos aspectos que merecen destacarse. Además de ser un estudio bastante amplio, ha comprendido aproximadamente 100 publicaciones, ha procurado demarcar los límites y alcances del test reformulado, como también analizar si los trabajos poseerían estructura y cuerpo suficiente para sacar conclusiones.

La propuesta Goodenough-Harris (Harris, 1991) al ser correlacionada con la original de Goodenough (1926) ha mostrado coeficientes alrededor del 0,86, en relación a la figura del hombre, común a las dos escalas. Estudiada con otras 14 medidas de inteligencia el coeficiente promedio fue alrededor de 0,49. Comparada al Wisc-R y al Stanford-Binet se ha constatado que hubo una tendencia a subestimar los puntajes de las personas, notablemente en los intervalos superiores de la inteligencia.

Uno de los hallazgos de Scott (1981) fue que, a pesar de ser pocos los estudios, los resultados han sugerido que los puntajes del test original de Goodenough (1926) se aproximarían más a los tests de inteligencia. Añádase a eso el hecho de que el test Goodenough-Harris (Harris, 1991) ha demostrado poca utilidad para predecir la realización académica, a diferencia del sistema de Goodenough.

Scott (1981) ha relatado también que en 24 de las 33 comparaciones entre los sexos encontradas en los estudios analizados no fue señalada diferencia significativa alguna. En las 9 que han informado diferencias significativas, las niñas han obtenido puntajes más altos que los niños en siete casos. Sin embargo, no fue observada ninguna característica que pudiera justificar ese resultado. A ese respecto, en el estudio de Sinha (1970) el resultado del análisis factorial ha sugerido que los niños han tenido más preocupaciones en relación a la proporción y que las niñas han estado más atentas a los detalles.

Entre otras conclusiones Scott (1981) ha afirmado que todo indica que el test de Goodenough-Harris es una medida estable y confiable, pero sería un predictor pobre del desempeño en los principales tests de inteligencia. Sin embargo, podría ser interesante si usado como *screening* para seleccionar personas con inteligencia abajo del promedio.

Investigaciones posteriores han informado que no obstante su uso difundido no fue posible comprobar que el Dibujo de la Figura Humana por el sistema Goodenough-Harris mediría la inteligencia como las Matrices Progresivas de Raven, la Escala de Inteligencia de Stanford-Binet-Binet, el Porteus Maze Test, el Wisc-R y la Escala de Inteligencia de Wechsler para niños, entre otras, parecen capaces de hacer (Abell, Von Briesen, & Watz, 1996; Harris, 1991; White, 1979).

Además de eso, investigaciones han señalados resultados poco estimulantes para el uso de ese sistema para estimar la inteligencia de niños (Aikman, Belter, & Finch, 1992; Kamphaus & Pleis, 1991). Más aún, algunos investigadores (Gresham, 1993; Motta, Little, & Tobin, 1993a; Motta, Little, & Tobin, 1993b) argumentaron que pese a que no se puede negar la validez de uso del Dibujo de la Figura Humana de Goodenough-Harris, otros tests han proporcionado resultados más válidos, lo que haría superfluo su uso. En compensación, otros autores (Bardos,

1993; por ejemplo) contra-argumentaron que algunos estudios realmente han hallado poca evidencia de validez para su uso, pero nuevas interpretaciones y sistemas para la evaluación de la Figura Humana estarían disponibles y los críticos parecen no tener eso en cuenta.

En el Brasil hay investigaciones estudiando el DFH en relación al nivel socioeconómico de preescolares (Van Kolck, 1981), diferencias entre sexos (Almeida, 1959), niños deficientes mentales (Carvalho, 1960), relaciones entre tests de inteligencia (Carvalho, 1960), relaciones con pruebas piagetianas (Sisto, 2000), desempeño escolar (Bandeira & Hutz, 1994), estandarización (Antipoff, 1931; Hutz & Antoniazzi 1995), validez y estandarización (Alves, 1981; Sisto, 2005), entre otras. En suramerica es también estudiado (Barros & Ison, 2002; por ejemplo).

Principalmente dos hechos son constantes en la literatura sobre el DFH. Uno de ellos se refiere a la validez para evaluar la inteligencia de las personas y la otra la diferencia entre los géneros, diferencia esa ya apuntada por Goodenough (1926). Con base en esas informaciones se ha elegido estudiar el aspecto referente a la validez interna de los ítems, más específicamente en relación al funcionamiento diferencial de los ítems en lo que atañe a los géneros.

El tema no es nuevo. Binet y Simon (1916) al estudiar niños de *status* socioeconómico más bajo que tenían un rendimiento peor en algunos ítems de su test, han planteado la posibilidad de que ellos podrían estar midiendo efectos de aprendizaje cultural y no de capacidad mental.

Es sabido que cuando un test tiene en cuenta los requisitos psicométricos de precisión y validez proporcionará medidas de personas con un margen de error muy pequeño. Los sesgos de los ítems pueden ser calificados teniendo problemas concernientes a la posibilidad de interpretación de los resultados del test, es decir, el grado en que el conjunto de ítems mide un rasgo o constructo. En ese sentido, en la teoría de los tests la probabilidad de que un examinado responda a un ítem correctamente se denomina probabilidad de éxito y los sesgos pueden ser estudiados comparando las probabilidades de éxito para diferentes subgrupos de una misma población. En otros términos, si el puntaje obtenido es función no sólo del nivel de los sujetos en la variable medida, sino también de otras características irrelevantes como pertenecer a diferentes grupos étnicos, culturales, entre otros, o en función de variables tales como sexo, o experiencia instruccional recibida, se trata de funcionamiento diferencial del ítem (DIF). Más específicamente, se refiere a una diferencia entre un grupo de referencia (personas del sexo masculino) y un grupo focal (personas del sexo femenino) en la probabilidad de acertar un ítem. Así, un ítem sesgado será aquel cuyas probabilidades de éxito son diferentes, pese a la igualdad de capacidad de las personas que respondieron a él. Por las implicaciones éticas, sociales y jurídicas involucradas en la utilización de tests que pueden subestimar sistemáticamente las capacidades de ciertos grupos, los estudios para neutralizar ese efecto son de importancia indiscutible.

Hay dos tipos de abordaje para la detección del sesgo de los tests. Una de ellas utiliza un criterio externo al test y la otra un criterio interno, normalmente los puntajes obtenidos en el test como un todo. En este estudio, el interés está en el sesgo interno que se refiere a las propiedades psicométricas de los ítems de los tests. En realidad se buscará responder si los ítems del test DFH poseen el mismo comportamiento estadístico (o equivalencia de medida) cuando son comparados a subgrupos de sujetos pertenecientes a la misma población. En el caso de que ese hecho sea observado, la conclusión es que no hay funcionamiento diferencial de los ítems (DIF); cuando la equivalencia no es constatada, se concluye por la presencia de DIF.

Hay diversos procedimientos para el estudio del funcionamiento diferencial del ítem (DIF), los cuales pueden clasificarse entre los que aplican la Teoría de Respuesta al Ítem (TRI) y los llamados de tablas de contingencia (Agueri, Zanelli, & Galibert, 2002; Aguerri, Galibert, Zanelli, & Attorresi, 2005; Benito & Ara, 1998; Fidalgo, 1996; Fidalgo, Mellenbergh, & Muñiz, 1998; Galibert, Aguerri, & Attorresi, 2000; Gómez & Hidalgo, 1997; Marañón, García, & Costas, 1999; Millsap & Everson, 1993). El modelo usado en este estudio fue el modelo logístico de Rasch (Rasch, 1960; Wright & Panchapakesan, 1969). Ese modelo puede incorporar gran parte de los trabajos precedentes sobre sesgo porque empieza con suposiciones similares de medida. Pero como sus procedimientos son extensiones racionales del modelo, el análisis del sesgo puede ser buscado adicionalmente y de manera más sistemática y más integrada de lo que ha sido hecho. En particular, ese procedimiento identifica los ítems que pueden conducir a una medida válida para toda la persona y pueden consecuentemente ser usados para detectar y corregir no solamente medidas sesgadas para cualquier grupo sino también para detectar una medida sesgada para el individuo.

Método

Participantes

Fueron investigados 2508 niños, siendo que el 49,76% (1248) era del sexo masculino y el 50,24% (1260) del sexo femenino. Esos niños frecuentaban desde la educación preescolar hasta en el curso cuarto de enseñanza primaria, en escuelas públicas (72,1%) y particulares (27,9%). Las escuelas se ubicaban en ocho diferentes ciudades del interior paulista. Las edades variaron entre 5 y 10 años, con un promedio de 8,1 (moda y mediana de 8,0) y una desviación típica de 1,30. La concentración más pequeña de niños se ha dado entre los que recibían educación preescolar, que fue poco más del 9% del total. En contraposición, la frecuencia más grande ha ocurrido en las edades de los 7 a los 10 años, comprendiendo más del 90% de las personas estudiadas.

Administración y criterios de corrección

La consigna para los niños fue que ellos dibujasen una persona lo más detallada posible, usando un lápiz y una hoja de

papel. Les fue permitido borrar el dibujo para correcciones. La administración fue colectiva a todos los niños de cada aula y se tardó un año lectivo para reunir todos los protocolos.

Los protocolos fueron corregidos basándose en los 51 detalles o criterios propuestos por Goodenough (1928). En todos los protocolos fue observada la presencia de cabeza, pierna y brazo mientras que dos otros criterios (perfil A y perfil B) no han tenido ocurrencia suficiente para análisis, así que esos ítems no han tenido variabilidad suficiente para cualquier análisis y fueron descartados. En total fueron usados 46 ítems para estudio y la amplitud posible del instrumento fue de 0-46. A los ítems les fue atribuido el valor de uno cuando el detalle estaba presente y de cero cuando el detalle estaba ausente.

Análisis

Fue usado el modelo Rasch por medio el programa Winsteps para los análisis. Primeramente, fueron estimadas las dificultades de los ítems para, a continuación, separar la distribución de la aptitud para cada una de las personas de los dos grupos. En otros términos, el modelo corrige la estimativa de las dificultades de los ítems por la distribución de la aptitud de la persona. En consecuencia, la dificultad estimada debería ser estadísticamente equivalente para los grupos en el caso de que ellos si distinguen por tan solamente la distribución de sus aptitudes.

Finalmente, el modelo calcula cuánto de DIF es añadido (positiva y negativamente) al ítem y, por la prueba t de Student, estima si la diferencia puede ser atribuida al acaso o no (Wright, Mead, & Draba, 1976). Frecuentemente un valor de t de más de 2 es considerado significativo. Pese a eso, Draba (1977) concordando con Bonferroni, considera que 2,4 es un buen índice cuando se está analizando más de 20 ítems como es el caso del presente estudio.

Resultados

Los estadísticos descriptivos han mostrado que la media del grupo fue de 17,52 ($SD=7,09$) y el error estándar de 0,14. Los puntajes obtenidos han variado entre 0-46 puntos y la distribución de los datos fue bastante parecida a una curva normal simétrica. La media del grupo femenino (18,09) fue mayor que la obtenida por el grupo masculino (16,96), siendo la diferencia entre ellos significativa ($t=-4,06$; $p=0,000$).

Para verificar el ajuste al modelo Rasch fueron analizados los valores de *infit* y *outfit*, cuyo valor esperado es de 1. Se consideran que los valores superiores a 1,5 indican un desajuste moderadamente alto y los superiores a 2,0 muy alto, de tal manera que perjudican gravemente las medidas (Wright & Linacre, 1998). Los resultados del análisis del ajuste de los ítems al modelo y de los participantes están en la Tabla 1.

Así que en relación al *infit*, todos los ítems se han ajustado al modelo como también las medias y las desviaciones típicas de los valores del *infit* y *outfit* han sido las esperadas cuando no hay divergencias substanciales

entre las previsiones del modelo y los datos empíricos. A su vez, en lo que atañe al *outfit* tres detalles no han presentado ajuste, lo que supone que respuestas inesperadas en relación al nivel de aptitud fueron dadas, dos de ellas bastante preocupantes. Sin embargo, mientras que el *infit* es más robusto el *outfit* es un indicador bastante sensible a los *outliers* (anómalos), puesto que los altos valores se deben a respuestas netamente absurdas. Hay que considerar, también, que el porcentaje de los sujetos que no se han ajustado al modelo fue muy bajo.

En la Tabla 2 se presentan las estadísticas descriptivas de los parámetros de los ítems y de las personas. La fiabilidad fue alta tanto para las personas cuanto para los ítems. La amplitud medida fue bastante alta (10,06).

Aplicadas las técnicas de detección del funcionamiento diferencial de los ítems se ha identificado que de los 46 ítems estudiados, 20 de ellos ha diferenciado los sexos. Las Tablas 3 y 4 presentan los ítems fáciles para los sexos femenino y masculino, respectivamente, la adición para cada sexo, los valores de los cambios observados, los valores de t y la dificultad del ítem. Como puede ser observado, nueve ítems fueron de fácil ejecución para las niñas y 11 para los niños.

Comparando esos resultados a los de Goodenough (1926) se ha observado que de los 11 ítems indicados por ella como característicos del sexo femenino, tan solamente dos favorecieron las niñas en función del modelo de análisis aquí utilizado, los cuales son, la nariz y detalles del ojo con cejas y pestañas. En relación a los siete ítems que Goodenough ha observado como característicos del sexo masculino, ningún de ellos fue constatado por el análisis hecho en este estudio. Una posible interpretación para eso es que las diferencias pueden estar en función del tipo de análisis e instrumento usados, así como también que algunos de los ítems observados no forman parte de su escala final para la evaluación cognitiva.

Conclusiones

Los tests psicológicos son parte fundamental de los procesos de evaluación que implican toma de decisiones y que establecen diferencias entre grupos. Si los ítems de un test presentan problemas de DIF, los puntajes para los grupos involucrados no son comparables y, por lo tanto, ellos no pueden ser interpretados de igual modo. La posible falta de equidad de

Tabla 1.
Parámetros de ajuste de los ítems y de las personas

Parámetro	<i>Infit</i> (Ítems)	<i>Outfit</i> (Ítems)	<i>Infit</i> (Personas)	<i>Outfit</i> (Personas)
Media	0,99	1,01	1,00	1,01
D.T.	0,12	0,37	0,26	0,99
Máximo	1,36	2,16	2,31	9,90
N y (%) > 1,5	0 (0,00)	1 (2,17)	68 (2,71)	27 (1,08)
N y (%) > 2,0	0 (0,00)	2 (4,34)	8 (0,08)	95 (3,79)

Tabla 2.
Parámetros de los puntajes de los ítems y de las personas

Parámetro	Ítems	Personas
Media	0,00	-0,78
D.T.	2,25	1,32
Error estándar	0,34	0,03
Máximo	4,68	3,66
Mínimo	-5,77	-6,40
Confiabilidad	1,00	0,87

Tabela 3.
Ítems más fáciles para las niñas, adiciones, cambio y valores de t

Ítem	Adición para las niñas	Adición para los niños	cambio	t	Dificultad
7b	-0,30	0,22	-0,52	-3,89	3,08
8a	-0,67	0,42	-1,08	-7,25	3,37
9a	-0,41	0,33	-0,74	-6,96	2,10
9b	-0,24	0,25	-0,49	-5,46	0,47
13	-0,34	0,46	-0,80	-7,05	1,16
14a	-0,20	0,19	-0,39	-4,27	0,98
14f	-0,13	0,16	-0,29	-2,57	1,60
16a	-0,34	0,38	-0,72	-7,81	0,20
16c	-0,18	0,21	-0,39	-3,63	0,91

Tabla 4.
Ítems más fáciles para los niños, adiciones, cambios y valores de *t*

Item	Adición para las niñas	Adición para los niños	Cambio	<i>t</i>	Dificultad
10a	0,16	-0,16	0,32	2,93	2,41
10b	0,18	-0,19	0,37	3,75	0,70
10c	0,32	-0,31	0,63	5,05	1,82
10d	0,34	-0,32	0,65	4,78	2,05
11b	0,48	-0,44	0,92	8,09	1,14
12a	0,54	-0,51	1,05	10,92	0,38
12b	0,23	-0,23	0,46	4,95	0,40
12e	0,17	-0,17	0,35	2,98	2,65
14c	0,17	-0,17	0,34	2,60	1,84
15a	0,34	-0,33	0,68	6,29	1,06
15b	0,27	-0,26	0,53	3,61	2,25

los instrumentos de medida ha convertido los estudios de DIF en parte esencial en la construcción de tests y de sus reevaluaciones.

Este estudio no tuvo el objetivo de identificar las causas del DIF, sino más bien verificar su existencia entre 46 detalles, según el sistema de Goodenough, usados como criterio para evaluar el DFH. Se trató de una investigación para detectar el DIF, o sea, determinar una posible diferencia entre las conductas de los ítems, comparando dos grupos.

Muñiz (1997) ya ha dicho que es posible afirmar que no hay ítems completamente sin sesgos. Así que el problema sería saber cual es la cantidad tolerable de sesgo en un test. De hecho, casi la mitad de los ítems examinados de los criterios de Goodenough presentaron DIF, lo que puede ser calificado como alto. Ese resultado puede ser considerado como un indicativo de invalidez del test, pues dificultaría la interpretación de los resultados; sería muy difícil saber si un puntaje de un niño está rebajado porque él ha dibujado un detalle más característico de las niñas en detrimento de otro más peculiar a los niños.

Así, de los resultados encontrados parece desprenderse la hipótesis de que algún tipo de factor relacionado con el sexo de los sujetos puede estar influyendo en la forma como ellos dibujan una persona. Por eso, a lo mejor se deba considerar seriamente la posibilidad de intentar construir escalas diferenciadas para cada uno de los sexos, con ítems característicos de cada sexo. Tal vez si se hicieran escalas que consideren ese hecho el DFH, según el sistema de Goodenough, ofrezca evidencias de validez más consistentes. No se debe dejar de tener en consideración que ese sistema es el que mejor se correlaciona con otros tests de inteligencia y desempeño académico (Scott, 1981).

Aunque sea difícil que el DFH ofrezca psicométricamente condiciones similares a otros tests de inteligencia, no se puede dejar de lado el hecho de que es un instrumento que posibilita una evaluación rápida, no es invasor y que facilita su administración en personas con problemas de varias naturalezas. En total, su utilidad como *screening* no debe ser descartada, pues parece ser muy adecuado para determinadas situaciones. En ese sentido, la intención de profundizar los análisis de ese

instrumento deberían ser retomados, pues parece que las críticas planteadas a él es posible que sean planteadas para otros tests de inteligencia, como aquellos que presentan una gran cantidad de varianza no explicada cuando se correlacionan con otros tests de inteligencia y bajas correlaciones con el desempeño académico.

Un hecho que quizá también deba ser mejor analizado en el DFH se refiere a no captar inteligencias más altas (Scott, 1981). No obstante, mismo que se confirme su poca discriminación para personas con ese nivel, los tests no tienen ni estiman la inteligencia en un gran rango y lo que es alta inteligencia en algunos tests no suele ser en otros.

Es posible concebir que las diferencias observadas en la conducta de un ítem en personas de distintos grupos se pueden deber no a una validez diferencial del instrumento para los distintos grupos sino más bien a la diferente precisión con la que se han estimado los parámetros en uno y otro grupo. Pese a eso, lo más frecuente es que la pertenencia a un grupo determinado puede enmascarar variables de gran significación para el constructo pretendidamente evaluado (Muñiz, 1997). A ese respecto, Goodenough (1926) ya había apuntado para detalles que serían característicos del sexo masculino y otros del sexo femenino. Comparando los ítems hallados por ella y en esta investigación se ha observado que muy pocos fueron los mismos. De ese modo, pese a la diferencia de procedimientos para detectar cuales serían más característicos de un grupo u otro, hay algo que necesita más investigación para una mejor comprensión de lo que significan esos sesgos hallados. A la vista de los resultados, a lo mejor convendría hacer más investigaciones para entender mejor ese test y, si posible, realizar inferencias más fidedignas y válidas a partir de él.

Referencias

- Abell, S. C., Von Briesen, P. D., & Watz, L. S. (1996). Intellectual evaluations of children using human figure drawings: An empirical investigation of two methods. *Journal of Clinical Psychology*, 52(1), 67-74.

- Agueri, M. E., Galibert, M. S., Zanelli M. L., & Attorresi, H. F. (2005). Detección errónea del funcionamiento diferencial del ítem. Una comparación de métodos. *Psicothema*, 17 (2), 350-355.
- Agueri, M. E., Zanelli M. L., & Galibert, M. S. (2002). Evaluación de un método empírico para detectar el funcionamiento diferencial del ítem. *Interdisciplinaria*, 19 (2), 185-213.
- Aikman, K. G., Belter, R. W., & Finch, A. J. (1992). Human figure drawings: Validity in assessing intellectual level and academic achievement. *Journal of Clinical Psychology*, 48(1), 114-120.
- Almeida, R. M. (1959). Um estudo do status mental em um grupo de crianças nordestinas em idade escolar. *Boletim de Psicologia*, 11 (38), 35-55.
- Alves, I. C. B. (1981). O teste Goodenough-Harris em pré-escolares paulistanos. *Boletim de Psicologia*, 80, 33, 40-52.
- Antipoff, H. (1931). O desenvolvimento mental da criança de Bello Horizonte. *Revista da Educação e Saúde Pública*, 17, 17-27.
- Bandeira, D. R., & Hutz, C. S. (1994). A contribuição dos testes DFH, Bender e Raven na predição do rendimento escolar na primeira série. *Psicologia: Teoria e Prática*, 10, 1, 59-72.
- Bardos, A. N. (1993). Human figure drawings: Abusing the abused. *School Psychology Quarterly*, 8(3), 177-181.
- Barros, M. C., & Ison, M. S. (2002). Conductas problemáticas infantiles: indicadores evolutivos y emocionales em el dibujo de la figura humana. *Revista Interamericana de Psicología*, 36(1-2), 289-298.
- Benito, J. G., & Ara, M. J. N. (1998). Impacto y funcionamiento diferencial de los ítems respecto al género en una prueba de aptitud numérica. *Psicothema*, 10(3), 685-696.
- Binet A., & Simon T. (1916). *The Development of Intelligence in Children*. Transl. ES Kite. Baltimore: Williams Wilkins
- Brown, W. R., & McGuire, J. M. (1976) Current psychological assessment practices. *Professional Psychology*, 7(4), 475-484.
- Carvalho, M. M. M. J. (1960). O desenho da figura humana como medida de inteligência e diagnóstico da personalidade em débeis mentais. *Boletim no.251, Psicologia*, 8, 29-44.
- Draba, R. E. (1977). The identification and interpretation of item bias. *Rasch Measurement Transactions*, MESA Memorandum no. 25. Recuperado em 13 de abril de 2004, de <http://www.rasch.org/rmt/rmt122m.htm>.
- Fidalgo, A. M., Mellenbergh, G. J., & Muñoz, J. (1998). Comparación del procedimiento Mantel-Haenszel frente a los modelos loglineales en la detección del funcionamiento diferencial de los ítems. *Psicothema*, 10(1), 209-218.
- Fidalgo, A.M. (1996). Funcionamiento diferencial de los ítems. In J. Muñoz (Ed.), *Psicometría* (pp. 371-455). Madrid, Spain: Universitas.
- Galibert, M. S., Aguerri, M. E., & Attorresi, H. F. (2000). Pesos óptimos de los ítems en la elaboración de los puntajes. *Revista Latinoamericana de Psicología*, 32(2), 79-90.
- Goh, D. S., Teslow, J., & Fuller, G. B. (1981). The practice of psychological assessment among school psychologists. *Professional Psychology*, 12(6), 696-706.
- Gómez, J., & Hidalgo, M. D. (1997). Evaluación del funcionamiento diferencial en ítems dicotómicos: Una revisión metodológica. *Anuario de Psicología*, 74, 3-32.
- Goodenough, F. L. (1926). *Measurement of intelligence by drawings*. World Book Company, The House of Applied Knowledge, New York.
- Gresham, F. M. (1993). "What's Wrong in This Picture?": Response to Motta et al.'s Review of Human Figure Drawings. *School Psychology Quarterly*, 8 (3), 182-86.
- Harris, D. B. (1991). *El test de Goodenough. Revisión, ampliación y actualización*. Espanha: Ediciones Paidós.
- Hutz, C. S., & Antoniazzi, A. S. (1995). O desenvolvimento do desenho da figura humana em crianças de 5 a 15 anos de idade: normas para sua avaliação. *Psicologia: Reflexão e Crítica*, 8(1), 3-18.
- Kamphaus, R. W., & Pleiss, K. L. (1991). Draw-A-Person techniques: Tests in search of a construct. *Journal of School Psychology*, 29(4), 395-401.
- Koppitz, E. M. (1968). *El dibujo de la figura humana en los niños*. Buenos Aires: Editorial Guadalupe.
- Louittit, C. M., & Browne, C. G. (1947). The use of psychometric instruments in psychological clinics. *Journal of Consulting Psychology*, 11, 49-54.
- Marañón, P. P., García, B. M. I., & Costas, C. S. L. (1999). Detección del funcionamiento diferencial de los ítems en una prueba de ciencias. *Psicothema*, 11(3), 691-697.
- Millsap, R. E., & Everson, H. T. (1993). Methodology Review: Statistical Approaches for Assessing Measurement Bias. *Applied Psychological Measurement*, 17 (4), 297-334.
- Motta, R. W., Little, S. G., & Tobin, M. I. (1993a). The use and abuse of human figure drawings. *School Psychology Quarterly*, 8 (3), 162-169.
- Motta, R. W., Little, S. G., & Tobin, M. I. (1993b). A picture is worth less than a thousand words: Response to reviewers. *School Psychology Quarterly*, 8(3), 197-199.
- Muñiz, J. (1997). *Introducción a la teoría de respuesta a los ítems*. Madrid: Psicología
- Naglieri, J. A. (1988). *Draw a Person: A quantitative scoring system. Manual*. The Psychological Corporation Harcourt Brace Jovanovich, Inc. Pirámide.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Nielson & Lydiche.
- Scott, L. H. (1981) Measuring intelligence with the Goodenough-Harris Drawing Test. *Psychological Bulletin*, 89, 1, 483-505.
- Sinha, M. (1970). A study of the Harris Revision of the Goodenough Draw-a-Man test. *British Journal of Educational Psychology*, 40, 221-222.
- Sisto F. F. (2005). *O Desenho da Figura Humana – Escala Sisto*. Vetor Editora Psicopedagógica Ltda.
- Sisto, F. F. (2000). Relationships of the Piagetian Cognitive development to Human Figure Drawing. *Journal of School Psychology*, 30 (4), 432 – 441.
- Van Kolck, O. L. (1981). *Técnicas de exame psicológico e suas aplicações no Brasil*. Petrópolis: Vozes.
- White, T.H. (1979). Correlations among the WISC-R, PIAT, and DAM. *Psychology in the Schools*, 16(4), 497-501.
- Wright, B. D., Mead R., & Draba R. (1976). Detecting and correcting test Item Bias with la Logistic Response Model. *MESA Research Memorandum*, no. 22. Mesa Psychometric Laboratory. Recuperado em 12 de abril de 2004, de <http://www.rasch.org/rmt/rmt122m.htm>.
- Wright, B.D., & Linacre, J.M. (1998). *WINSTEPS: A Rasch computer program*. Chicago: MESA Press.
- Wright, B.D., & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement*, 29, 23-37.

Received 25/07/2006
Accepted 21 /11/2006

Fermino Fernandes Sisto é doutor pela Universidad Complutense de Madrid, Livre – docente pela Unicamp e docente do curso de Psicologia e do Programa de Estudos Pós-graduados em Psicologia, da Universidade São Francisco, campus Itatiba-SP. Bolsista produtividade do CNPq.