Using Four-Parameter Item Response Theory to model Human Figure Drawings

Ricardo Primi[1]
Universidade São Francisco, Campus Swift, Campinas-SP, Brasil
Tatiana de Cassia Nakano, Solange Muglia Wechsler
Pontifícia Universidade Católica de Campinas, Campinas-SP, Brasil

ABSTRACT

This paper studied the application of an extended version of the Item Response Theory, the 4-parameter model (4PL), in the item analysis of Human Figure Drawing (HFD). HFD score drawing might serve as indicators of cognitive development. This model incorporates an upper asymptotic parameter (parameter *d*) admitting the possibility that children with high capacity have a probability less than 1 to draw a certain HFD detail (item). This is often observed in HFD. We performed IRT model three times using 1, 2 and 4 parameter models and compared their model fit indexes. The latent trait correlations estimated by these three models were very high (r=0.98), suggesting that children's abilities did not change substantially when using the 4-parameter model. It is pointed out a limitation in the correct way of modeling test item dimensionality considering that there is a hierarchical structure among items.

Keywords: Four parameter Item Response Theory; optimal scoring; psychometrics; intelligence assessment.

RESUMO – Usando a Teoria de Resposta ao Item de quatro parâmetros no Desenho da Figura Humana

Esse artigo estudou a aplicação de uma versão estendida da Teoria de Resposta de Item, o modelo de 4 parâmetros (4PL), na análise dos itens do Desenho de Figura Humana (DFH). O DFH pontua detalhes do desenho como indicadores do desenvolvimento cognitivo. Esse modelo incorpora um parâmetro assintótico superior (parâmetro *d*) admitindo a possibilidade de que crianças com alta capacidade tenham probabilidade menor que 1 de desenhar determinado detalhe (item) do DFH. Isso é um evento comum no DFH. Executamos o modelo de TRI três vezes, usando modelos de 1, 2 e 4 parâmetros e comparamos seus índices de ajuste. As correlações dos traços latentes estimados por esses três modelos são muito altas (r=0,98), sugerindo que as habilidades das crianças não mudaram substancialmente ao usarmos o modelo de quatro parâmetros. Aponta-se uma limitação na maneira correta de modelar a dimensionalidade dos itens do teste considerando que há uma estrutura hierárquica dos itens.

Palavras-chave: teoria de resposta ao item de quatro parâmetros; escores ótimos; psicometria; avaliação da inteligência.

RESUMEN – Uso de la Teoría de la Respuesta al Ítem de cuatro parámetros en el Dibujo de la Figura Humana

Este artículo estudió la aplicación de una versión extendida de la Teoría de la Respuesta al Ítem, el modelo de 4 parámetros (4PL), en el análisis de los ítems del Dibujo de la Figura Humana (DFH). El DFH puntúa los detalles del diseño como indicadores del desarrollo cognitivo. Este modelo incorpora un parámetro asintótico superior (parámetro *d*) admitiendo la posibilidad de que niños con alta capacidad tengan probabilidad menor que 1 de dibujar determinado detalle (ítem) del DFH. Esto es un evento común en el DFH. Se ejecutó el modelo TRI tres veces, usando modelos de 1, 2 y 4 parámetros y se comparó sus índices de ajuste. Las correlaciones de los rasgos latentes estimados por estos tres modelos son muy altas (r=0,98), sugiriendo que las habilidades de los niños no cambiaron sustancialmente al usar el modelo de 4 parámetros. Se apunta una limitación en la manera correcta de modelar la dimensionalidad de los ítems de la prueba, considerando que hay una estructura jerárquica de los ítems.

Palabras clave: Teoría de la respuesta al ítem de 4 parámetros; escores óptimos; psicometría; evaluación de la inteligencia.

Ever since the 19th century, research has revealed empirical interest in children drawing abilities. In 1926, Goodenough developed a test to assess children's intellectual development from human figure drawings (HFD). The method is based on findings of that general developmental stages exist, and that they impact drawings performed by children. Since then this method has been widely employed (Cronin, Gross, & Hayne, 2017). At the age of four, for example, authors argue that while some children can only make scribbles, other produce drawings in form of tadpoles, while others elaborate human figures with head, torso and four separated

[1] Endereço para correspondência: Universidade São Francisco. Rua Waldemar César da Silveira, 105, 13045-510, Campinas, SP. *E-mail:* rprimi@mac.com

members. Such individual differences in children figure drawings attributes have led some researchers to argue that they could be used as indicators of cognitive development. Drawings from older children contain more details surrounding the concept of the human body, being increasingly more realistic. Thus, these features indicate that children's drawings reveal the development of knowledge about human body concept, which could be viewed as an indicator of crystallized intelligence (Imuta, Scarf, Pharo, & Hayne, 2013; Wechsler, 1998).

After Goodenough (1926), some methods were produced in which authors reviewed her initial proposal. Instances are the systems developed by Machover (1949), Hammer (1958), Koppitz (1984), Harris (1965), Naglieri (1988) and Reynolds and Hickman (2004). HFD is a simple non-verbal task, easy to administer and score when compared to traditional measures of intelligence. Because of that, clinical and educational psychologists use HFD for screening or diagnosing cognitive abilities in children (Dans-Lopez & Tarroja, 2010). Other uses have also been reported in the literature that involved the assessment of emotional and social functioning in children (Zielona-Jenek, 2013).

Therefore, assessment models follow two different approaches: (a) cognitive assessment scoring the amount of drawing details such as presence or absence of specific body parts size, proportion, differentiated details between the male and female figure, and (b) emotional problems assessment based on global assessment of the drawing with attributes such as quality, omission of important details, presence of distortion or aberrations (for instance monsters, non-human figures), whose occurrence could indicate atypical emotional development.

In Brazil, HFD has been studied for decades as a tool for cognitive as well as projective assessment (Alves, 1981, 2015; Hutz & Antoniazzi, 1995; Silva, Pasa, Castoldi & Spessatto, 2010; Suehiro, Benfica, & Cardim, 2016). Wechsler (2003) proposed a system for cognitive assessment of children from 5 to 12 years old. It requests children to perform two drawings, a female and male figure. The system contains 58 drawing's details (some common, some specific to each female vs male figure) that are scored 1 if it is present and 0 if it is absent. It also classifies expected, common, unusual and exceptional items. According to the author, psychologists should consider expected items as developmental markers of typical age groups. Their absence would imply immaturity, neurological or emotional problems. Another important feature is the asymmetry between the developmental patterns exhibited by girls and boys, which should be considered in the assessment (Wechsler, Prado, Oliveira, & Mazzarino, 2011).

Research on HFD has focused on external validity, by exploring the correlation to traditional tests of intelligence. However, few studies applied more modern psychometric methods in the analysis of HFD. Going back to the Goodenough's (1926) work, we found that she validated items by inspecting their characteristic curves, considering the probability of observing a detail in the drawings (for example, head) as a function of age. Nevertheless, by that time, IRT models were not well developed. IRT models the probability of a correct response - in the context of the drawings, the probability of a given detail of the drawing – $P(\theta)=1$ as a function of a latent trait θ and item attributes.

Campbell and Bond (2017) applied Rasch analysis to test if Goodenough's original 51 items were unidimensional, an important assumption in any IRT model. They selected items to construct a prototype human figure drawing continuum that could be used to explore development. They also investigated if other drawings add information over the classical draw-a-man task. Sisto (2005) also used Rasch analysis trying to test the unidimensionality of Goodenough's items. He found that more than one dimension is required to account for the covariance between items. Similarly, Flores-Mendoza, Abad and Lelé (2005) used 2-parameter and Samejima's graded response model in the analysis of Wechsler's model to accommodate local dependency that exist between some items. For instance, they grouped presence of nose, nose in two dimensions and complete nose structure in a polytomic item scoring 0, 1, 2 and 3. All these studies show that items vary in the extent they measure a general versus specific factors.

The present study applies an IRT modeling that has not yet been investigated with HFD. We apply a 4-parameter item response model (4PL), which is believed to be more flexible to represent item characteristic curves patterns that could be observed in HFD items. Traditional 3-parameter IRT models contain parameters for item differences in difficulty (b), discrimination (a) and lower asymptote (c) - that is, correct answers for low ability subjects (also called pseudo guessing). The 4PL model adds an upper asymptote (d) for modeling items that do not reach a peak equal to one. HFD items certainly vary in difficulty and discrimination. Lower asymptote is supposed to be zero since the HFD comprises a response contrition task. However, it could be the case that some items would code details that do not always peak at $P(\theta)=1$, as traditional 2 and 3 IRT models assume. For instance, consider feminine clothing details, chin and forehead; it can be possible that some children even with high cognitive development (high θ) do not include these details in their drawings. Therefore, HFD

items could vary in upper asymptote. The 4-parameter model is written as follows:

$$P(u = 1 \mid \theta_j, a_i, b_i, c_i, d_i) = c_i + (d_i - c_i)\frac{\exp(a_i(\theta_j - b_i)}{1 + \exp(a_i\,(\theta_j - b_i))}$$

In this model, $P(\theta_j)=1$ is the probability of a person j scoring 1 in item i as a function of a latent latent trait θ_j. Each item i is characterized by its lower asymptote c_i, upper asymptote d_i, discrimination a_i and difficulty b_i. This study explores if there is evidence of upper asymptote less than one by applying the 4-parameter IRT and comparing it with traditional 1-PL and 2-PL model. It is expected that 4-PL model will be more appropriate to HFD data.

Method

Participants

The sample consisted of 3.144 participants, aged from 5 to 11 years (M=8.04, SD=1.79), of which 1.624 were female. Of these, 1.702 were students from private school, 1.422 from public schools, and 20 from 15 different Brazilian non-governamental organizations located in the states of São Paulo, Minas Gerais, Mato Grosso, Bahia, Distrito Federal, Sergipe, Amazonas, and Santa Catarina.

Instruments

Human Figure Drawing Test (Wechsler, 2003). The test asks children to draw a female and a male figure in two separated sheets of paper. The drawings are scored based on the presence or absence of 58 details in each drawing. It comprises 43 items common to male and female drawings, and 15 specific items to each drawing. The test can be administered in children from 5 to 12 years old, and each item is scored 1 for presence and 0 for absence. Total raw score for each drawing is computed and then transformed into percentile scores comparing children with a normative sample of children with same age and gender.

Also available is a qualitative analysis to check how many expected (items that are observed in 86 and 100% of drawings), common (51 and 85%), uncommon (16 to 50%) and exceptional (1 to 15%) items are present. Reliability and validity evidence is reported in instrument's manual (Wechsler, 2003) and studies conducted by other researchers (Bandeira, Costa, & Arteche, 2008; Flores-Mendoza et al., 2005; Wechsler, 1998; Wechsler & Schelini, 2002). Recently, the possibility of expanding the system to assess creativity (Oliveira & Wechsler, 2016) and emotional characteristics (Wechsler et al., 2011) have also been explored.

Data analysis

The main strategy for data analysis was to explore whether 4-parameter IRT modeling would be a better method to capture all available information on developmental patterns on children human figure drawings. We followed four steps: (a) we first ran an graphical item analysis of empirical item characteristic curves to visually check if there is evidence of upper asymptote less than 1 on the items; (b) we then ran IRT modeling three times using 1, 2 and 4 parameters models and compared more formally their model fit indices; (c) we examined d_i parameters to check if there is a substantial number that are lower than one; and (d) we finally compared item and person parameters estimated from these models. The main question was whether item parameters and person latent factor scores differ between models. In case they differ, we intended to explore why and to what extent theses scores will tell different stories about people in comparison with 2-parameter model and classical total score. Our overall question was if 4-parameter modeling could provide different information about children being potentially more valid scores. R packages mirt (Chalmers, 2012), *psych* (Revelle, 2017) and *tidyverse* (Wickham, 2017) were used in the analysis.

Results

Graphical item analysis

In the first set of exploratory analysis of empirical item characteristic curves, we observed that 35 out of 116 items indeed had the upper part of their curves not peaking at $P(\theta)=1$ (17 from female and 18 from male figure drawings). Figure 1 shows four examples from female figure drawings. Figure 1's x-axis represents groups of children with same total scores, that is, children with equal levels of ability (total score is a proxy for the latent scale score). From left to right the scores range from 0 to 58 (maximum possible). Y-axis shows the proportion of correct responses for each score group. Empirical probabilities were smoothed to capture the general pattern. Item 10e shows a typical pattern of increasing probabilities as ability increases approaching 1. The other three items (8b, 13d and 17c) show a similar pattern, never reaching 1, but peaking around .5 to .8. Main IRT models (1, 2 and 3 parameters) predict that, as ability increases, the item probability will be 1. This seems not to be true specially for items 8b, 13d and 17c, as well as others that reveal a similar pattern.

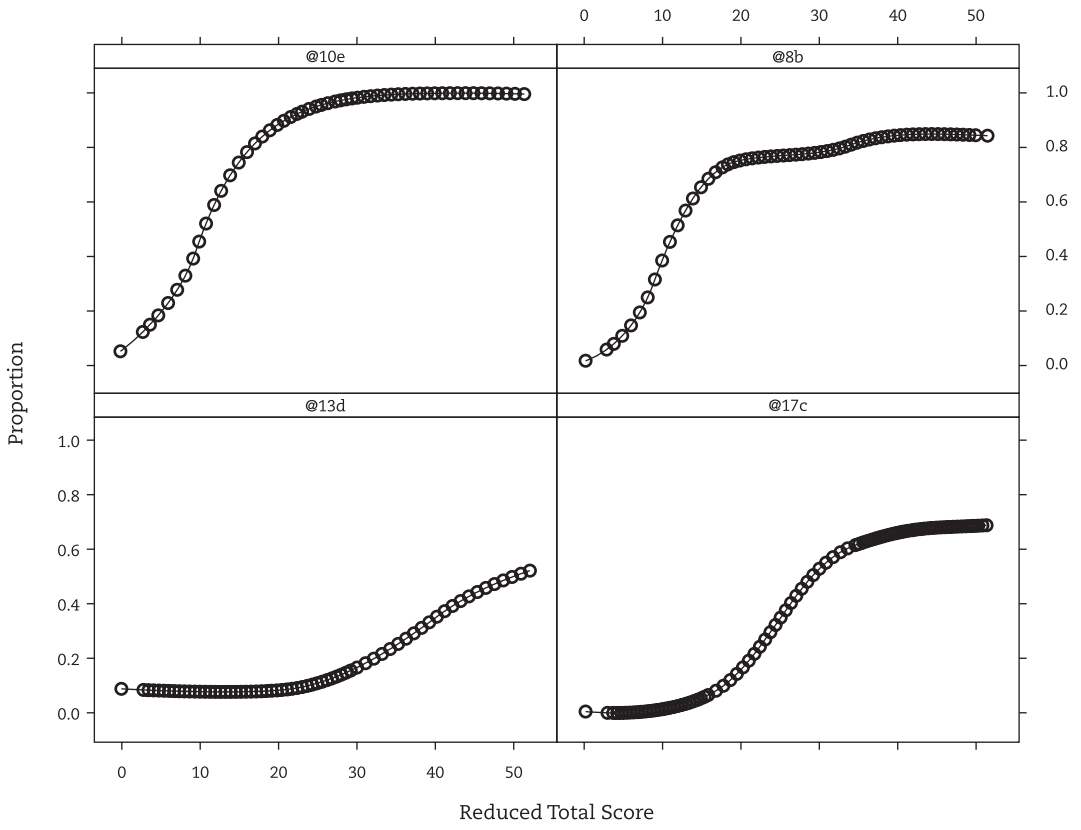

Figure 1. Examples of empirical Item Characteristic Curves with upper asymptote d_j less than one (items 8b, 13d, 17c) and a typical item (10e)

IRT modeling and model fit comparisons

The next step was to calibrate item parameters and estimate children latent scores. Before proceeding with IRT modeling, we tested if there is one principal dimension explaining inter-item correlations, that is, we tested the condition of unidimensionality required for IRT analysis. We ran an exploratory bi-factor analysis of the tetrachoric inter-item correlations matrix. We used omega function of package (Revelle, 2017) and guidelines of Rodriguez, Reise, and Haviland (2016) and Reise, Moore, and Haviland (2010). Scree plot and parallel analysis showed a general factor with the $\lambda_1/\lambda_2=4.44$, plus also five non-negligible group factors. Eighty-nine items (76%) had a loading greater than .29 on the general factor. The importance of this general factor, as measured by the Explained Common Variance, was ECV=.44. These indices suggest a more complex multidimensional structure underlying these items. An inspection of the loading matrix showed that group factors tended to cluster items referring to the same body parts (head, neck, hands and feet). We finally compared bi-factor loadings on the general factor – that accounted for the multidimensional structure with five group factors – with full-information loading of a unidimensional model performed by MIRT (Chalmers, 2012). The factor congruence coefficient was .97, indicating that the meaning of the IRT general factor is the same as the meaning of a general factor in a more complex multidimensional structure modeled by a bi-factor model. This indicates that the data satisfied the condition of essential unidimensionality, and that group factors are not distorting item parameters estimates of a unidimensional IRT model.

Examining *d* parameters

In the next step, we calibrated items using 1, 2 and 4 parameter models using MIRT package (Chalmers, 2012). Fit indices comparing these three models are presented in Table 1. As can be seen, 4-parameter and 2-parameter model provided a better fit to the data, as would be expected. The 4-parameter model was slightly better than the 2-parameter model. We then investigated if there is a substantial number of items with $d_j<1$. Figure 2 shows a histogram of the *d* parameters. Most of the items had *d* close to 1, but also some items had their values spread bellow 1.

Table 1
Fit indices of the IRT models.

Model	M2	*df*	*p*	RMSEA	SRMSR	TLI	CFI
1-PL	117919.77	6669	0	0.073	0.097	0.785	0.785
2-PL	102959.56	6554	0	0.068	0.080	0.810	0.813
4-PL	92649.78	6438	0	0.065	0.075	0.827	0.833

Note. Fit indices were: RMSEA=Root Mean Square Error of Approximation; SRMSR=Standardized Root Mean Square Residual; CFI=Comparative Fit Index; TLI=Tucker-Lewis index

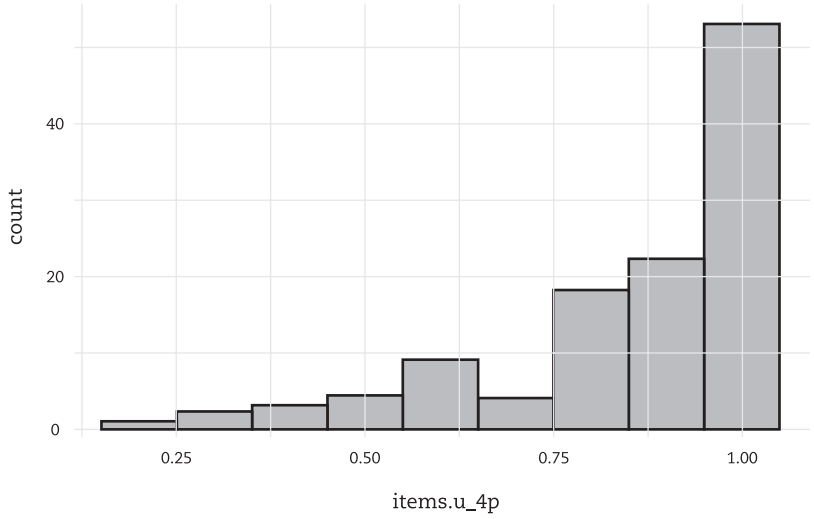

Figure 2. Histogram of the d_i upper asymptote parameters

Comparison of item and person parameters

We then compared item parameters obtained from the three models. Table 2 shows the observed correlations among parameters. The *b* parameters were very similar across models. Discrimination a parameters were moderately related in the 2- and 4-parameter models. Figure 3 shows the correspondence of item difficulties from 2- and 4-parameter models. Points are colored by the values of parameters *d*. It can be seen that the unpaired match tends to have an upper asymptote *d* less than 1.

Table 2
Correlations among item parameters in 1-, 2- and 4-parameter models

	items.a_2p	items.b_2p	items.a_4p	items.b_4p	items.u_4p
items.b_1p	-.13	.91	-.40	.90	-.40
items.a_2p		-.24	.55	-	.37
items.b_2p			-.49	.82	-.47
items.a_4p				-.45	-
items.b_4p					-

The most important question we asked is if different models would tell a different story about examine abilities. Do the latent score estimates differ substantially between models? Table 3 shows the correlations among latent scores estimated by Expected a Posteriori method (EAP) for 1-parameter (F1_1p), 2-parameter (F1_1p), 4-parameter models (F1_4p), and classical total scores (S1). Surprisingly, despite the between-model differences, including the unity weighted classical test theory scoring method, correlations between estimates provided by these models were all higher than $r \geq .98$ indicating that these scores tell the same underlying story about children abilities. Figure 4 shows the scatter-plot of latent scores estimated from the 2- and the 4-parameter models. Although generally similar, the mismatch occurs at the high end of the latent scale. Why does this happen?

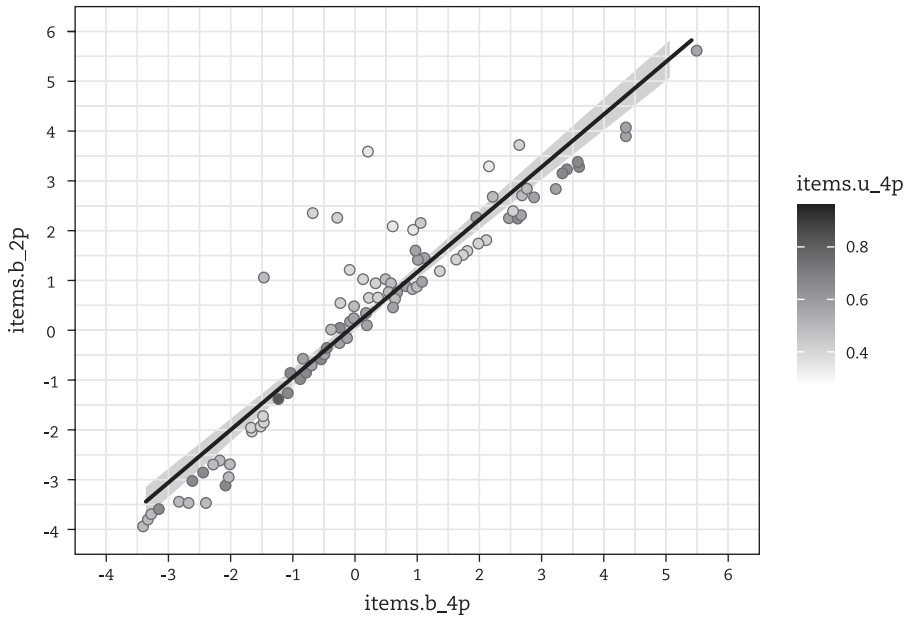

Figure 3. Correspondence of item difficulties from 2 and 4 parameter models

Table 3
Correlations of children's latent scores under 2- and 4- parameter models

	F1_1p	F1_2p	F1_4p
S1	.99	.98	.98
F1_1p		.99	.98
F1_2p			.99

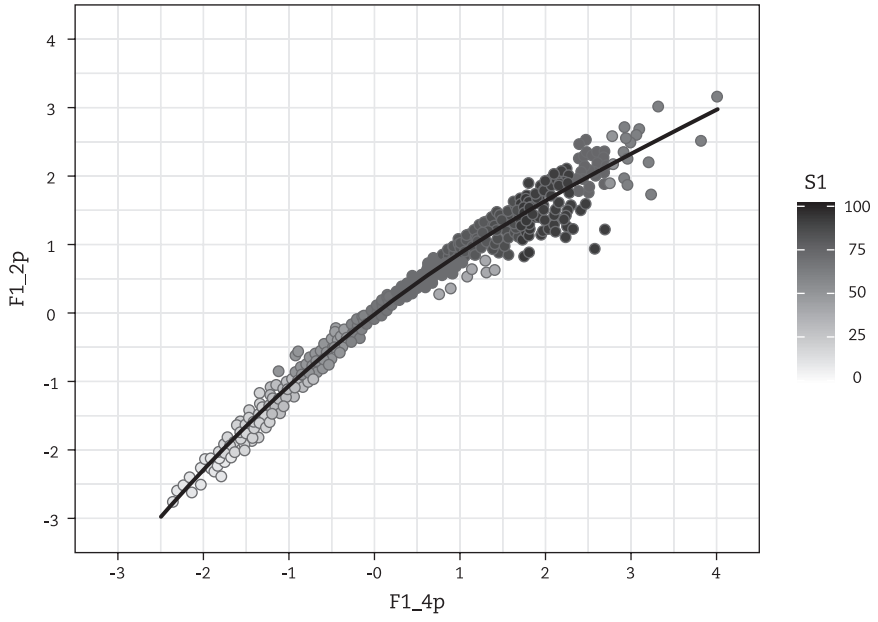

Figure 4. Scatterplot of children latent scores from 4- (X-Axis) and 2- (Y-axis) parameter models

How response patterns are scored on 3- and 4-parameter models?

To understand these differences, we need to review IRT equations that show the relationships of response patterns, item parameters and latent scores (the R code for this section is available at http://www.labape.com.br/rprimi/R/3pl_4pl_simulation.html. Lord (1980) shows that optimum item scoring weights is given by:

$$w_i(\theta_j) = \frac{P_i'(\theta_j)}{P_i(\theta_j)Q_i(\theta_j)}$$

Where w_i is the item i optimum weight, $P_i'(\theta_j)$ is the first derivative of the Item Characteristic Curve (ICC) of item i that depends on the model used, $P_i(\theta)$ is the probability of a person j scoring 1 in item i as a function of a latent latent trait θ_j given by the model being considered, and $Q_i(\theta_j)=1-P_i(\theta_j)$, that is, the probability of error – scoring 0 in item i as a function of a latent latent trait θ_j. The first derivative indicates the change in the probability of a correct response as a function of changes in θ (more precisely the derivative is the Y instantaneous rate of change of each point on X axis). This value changes as a function of theta, and it will be higher – indicating a greater change in probability – in the region closer to item difficulty b_i. This region is where the item is more discriminating, that is, where it occurs a greater rate of change in the probabilities as theta increases. It is intuitive that the item weight should be higher where the item is more discriminating. This pattern is reflected in this formula as the weight is directly proportional to a_i as will be seen in the next formulas.

This formula of optimum scoring weights helps to understand how each item is weighted when calculating a latent score for an examinee. Magis (2013) presents the first derivative of the 4-parameter model:

$$P'(\theta) = \frac{a_i}{d_i - c_i}[P(\theta) - c_i][d_i - P(\theta)]$$

Substituting $P'(\theta)$ on the formula for w_i (θ) we get the general formula for the optimum score weights for any IRT model (1-, 2-, 3- 4-parameter models):

$$w_i(\theta) = \frac{a_i(c_i - P(\theta))(d_i - P(\theta))}{d_i - c_i}$$

This formula shows that the item scoring weight w_i is a complex function of the item parameters (a_i, c_i and d_i) as well as θ via $P(\theta)$. But this complexity depends on the IRT model used. If we consider the 1-parameter model where $a_i=1$, $c_i=0$ and $d_i=1$, this formula simplifies to $w_i=1$. This reflects the known fact that in Rasch models, the total score - that is, unit weighted sum score - is a sufficient statistic to estimate theta. This means that the total score has all the information needed to estimate theta. This also implies that different response patterns (1 and 0 on different items), but with the same total score, will lead to the same estimated value of theta (De Ayala, 2009).

If we consider the 2-parameter model where $c_j=0$ and $d_j=1$, this formula simplifies to $w_i=a_i$. Therefore, item scores are weighted by item's discrimination in estimating theta scores. This is also a sufficient statistic. It also implies that different response patterns with the same total score can lead to different values of theta to the extent that a_j varies among items. If two examinees have the same total score but one gets his right answers on more discriminating items, then he will get a higher theta.

If we consider 3- and 4-parameter models, then these weights become more complex because now they are a function of θ. This dependency is counter intuitive. The core question we ask while using scoring weight is: How do I weight items to calculate theta? The answer in 3- and 4- parameter models is: That depends on the theta of the examinee. The counter intuitive idea is that while we want to know a weight to score items and to discover theta, we depend on examinee's theta, which we do not know yet, to discover the weight. This fact relates to the concept of sufficient statistics. That is, even if we have a response pattern and the respective item parameters, we are still unaware of how the items will be weighted to estimate theta. It will depend on if the examinee has a lower or higher ability. This is a fundamental aspect of 3 and 4-parameter models that explains differences we found when we compared estimated theta for 2- and 4-parameter modes. In order to clarify this fact, we explored this concept with simulation and data visualization. Let us consider a simple example with a 10 items test modeled with 3-parameter model with the following parameters:

$$\mathbf{a}=[0.8,0.8,0.8,0.8,1.2,1.2,1.2,1.2,1.2]$$
$$\mathbf{b}=[-2.4,-2.0,-2.0,-1.0,0,1.2,2.0,2.0,2.2,2.4]$$
$$\mathbf{c}=[0.30,0.23,0.08,0.21,0.12,0.11,0.34,0.21,0.09,0.15]$$

Let us consider two examinees A and B with following response patterns:

$$\mathbf{P_a}=[1,1,1,1,0,0,0,0,0,0]$$
$$\mathbf{P_b}=[0,0,0,0,0,0,1,1,1,1]$$

We can see that these two examinees scored 4, which indicates they both have low ability. Nevertheless, whereas examinee A got the easy items right, examinee

B responded correctly to the difficult items. What would be their latent scores estimates? Examinee A will be θ_a=-.54 and examinee B will be θ_b=-4.0! So, counter intuitively examinee B that got difficult items correctly responded got a lower score (note that these items were also more discriminating). Why does this happen? Figure 5 shows a matrix of weights of the 10 items (x--axis) by the ability level (y-axis). These weights were calculated using the general formula of optimum weights. Consider the upper part of the matrix where we can see item weights for examinees of lower ability (θ<0). We can see that difficult items got their weights equal to zero. That happens because, as Lord (1980) explains, "when low-ability examinees guess at random on difficult items, this produces a random result that would impair effective measurement if incorporated into the examinee's score; hence the need for a near--zero scoring weight" (p. 23). Since 3-parameter model assumes that correct responses on difficult items from lower ability examinees is a product of random guessing, it will penalize the latent score with a low weight (zero in its case) for difficult items. It is also interesting to note, in Figure 5, that maximum weight of an item is equal to its discriminating index a_j.

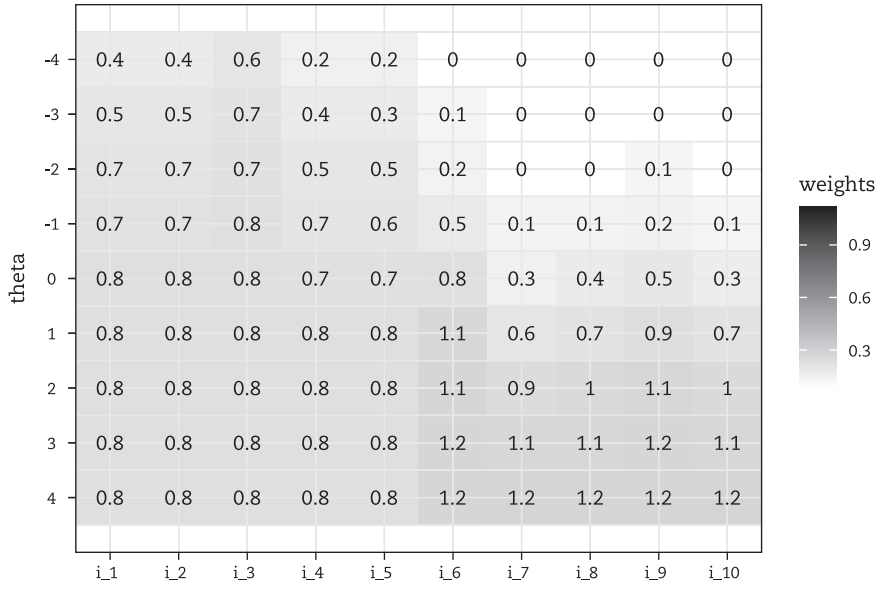

Figure 5. Scoring weights for Simulation 1 under 3-parameter model

What happens with item weights with the 4-parameter model? Let us modify our example with 10 items test modeled with 4-parameter model. The *a* and *b* parameter vectors keep the same values. The *c* parameters are all zero, consistent to the modeling we used in DHF and *d* parameters are:

$$\mathbf{d}=[.6,.6,.6,1,1,1,1,1,.6,.6]$$

This vector has two aspects: (a) an examinee could start the test unmotivated, so he or she could miss some items from the beginning of the test. Therefore, easier items that appear on the beginning of the test had *ds*=.6 accounting for expectation of wrong responses even for high ability examinees; (b) some items could identify peculiar human figure details, so that only a few individuals represents in their drawings even if they have high ability. Therefore, this is represented with difficult items with *ds*=.6. Let us consider two examinees, A and B, with following response patterns:

$$\mathbf{P_a}=[1,1,1,1,1,1,1,1,1,0]$$
$$\mathbf{P_b}=[0,0,1,1,1,1,1,1,1,0]$$

We can see that both examinees have high abilities because they get 9 and 7 items right. Examinee A is two points higher than examinee B so intuitively he should get a higher score. But their estimated theta will be the same θ_a=4! Here we see a similar pattern as discussed in the case of 3-parameter model but now in the high end of the scale. Figure 6 shows the same matrix of scoring weights for this second example. Consider the lower part of the matrix where we can see item weights for examinees of high ability (θ>0). We can see that easy items got their weights equal to zero. Therefore, although examinee B had lower total score, his misses occurred on the

easy items. Since the model expect slips in the easy items for examinees of high abilities these wrong responses doesn't count to lower subjects B latent score.

Therefore, coming back to the question why we saw different values of theta from 2-parameter model compared with correspondent 4-parameter model, we can see in Figure 4 that, for high ability children, for a same 2-parameter score, the correspondent 4-parameter score is more spread towards high ability. This is consistent with the simulations that wrong responses in items with $d<1$ do not lower the estimated theta, especially calculated from the easier items.

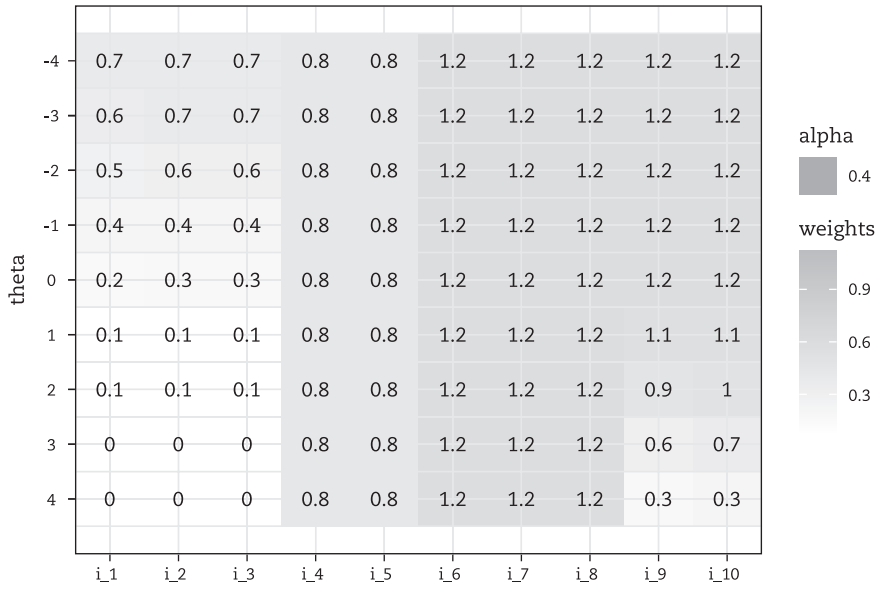

Figure 6. Scoring weights for Simulation 2 under 4-parameter model

Discussion

This paper tested whether 4-parameter IRT would be a better method to model items in Human Figure Drawings. This model adds an upper asymptote *d* accounting for the fact that some items could not reach probability of 1 in high levels of the latent trait. This could be true for some items referring to rare details on the drawings. Indeed, we found evidence that items vary in their *d* parameters, and that using a 4-parameter IRT model provided a better fit to the data as compared to one and two parameter models (we did not model c because the HFD is a task of constructed responses, the reason why a guessing parameter does not make sense in this case).

Despite the best fit of the 4-parameter, the correlations between latent trait estimates resulting from different models were very high ($r \geq .98$), then suggesting the final history about children abilities will not change substantially if we use a 4-parameter instead of a 2-parameter model latent scores. Even the classical total score is practically identical to the IRT-derived latent trait scores. Our conclusion is not supportive of use of this more complex IRT model since their final information about the subjects is like the ones obtained from simpler models. However, eventually, for a small group of children, scores from 4p to 2p will be different. These differences need to be considered in the context of error of measurement and the practical implications for the subjects tested. This result is similar to other studies that attempted to compare latent scores from different IRT models. See, for instance, Embretson and Reise (2000).

One important finding is that items in the Human Figure Drawings test have a more complex internal structure than a simple unidimensional model would predict. We found not only a large general factor, but also group factors among items, what have also been reported on previous factor analytic investigations of human figure drawings (Campbell & Bond, 2017; Flores-Mendoza, Abad & Lelé, 2005; Sisto, 2005). However, still no consensus exists concerning multidimensional structures of human figure drawing items and the usefulness and incremental information of specific factors beyond the general factor. This suggests an area for future research. It will be important to define these group factors and test their utility for children assessment (criterion validity). Human Figure Drawings are complex because items are constructed responses and organized hierarchically. For instance, the chance you draw a face with earrings depends on drawing ears in the first place. Hence, earrings

are nested within ears. Probably, items are nested inside macro body parts. Our bi-factor results suggest that head, neck, hands and feet are candidates for this macro-regions, as items apparently reflect elaboration on these parts of the body. It might be important to examine the internal structure of items using more complex factor analysis methods such as multilevel factor analysis (Muthén, 1991).

One additional contribution of this paper is the study on how scores are computed under 3PL and 4PL for different response patterns. We explored how the assumptions of 3PL and 4PL explain non-intuitive results that emerge when scoring response patterns. In IRT world, 3PL and 4PL models are the ones where we observe most changes when we compare their estimates to traditional total score from classical test theory. The main reason for this difference is the modeling of guessing behavior on difficult items (3PL) and carelessness on easy items (4PL). One important point is that these are strong assumptions that not always hold. See, for instance, Chiu and Camilli (2013), Andrich, Marais, and Humphry (2012). It will be interesting to study the implications of the use of 3PL model in scoring student responses on the Brazilian National Exam of High Education (ENEM). The main question would be if assumptions of the guessing behavior that are modeled in 3PL are feasible in the case of ENEM.

References

Alves, I.C.B. (1981). O teste Goodenough-Harris em pré-escolares paulistanos [Goodenought-Harris test in paulistanos preschoolers]. *Boletim de Psicologia, 80*(33), 40-52.

Alves, I.C.B. (2015). O desenho da figura humana para avaliação da inteligência de adultos analfabetos [The Human Figure Drawing for intellectual assessment of illiterate adults]. *Boletim – Academia Paulista de Psicologia, 35*(88), 75-92. Retrieved from http://pepsic.bvsalud.org/pdf/bapp/v35n88/v35n88a06.pdf

Andrich, D., Marais, I., & Humphry, S. (2012). Using a theorem by Andersen and the Dichotomous Rasch Model to assess the presence of random guessing in multiple choice items. *Journal of Educational and Behavioral Statistics, 37*(3), 417-442. doi:10.3102/1076998611411914

Bandeira, D. R., Costa, A., & Arteche, A. (2008). Estudo de validade do DFH como medida de desenvolvimento cognitivo infantil [The Draw- a Person test as a valid measure of children's cognitive development]. *Psicologia: Reflexão e Crítica, 21*(2),332-337. doi:10.1590/S0102-79722008000200020.

Campbell, C., & Bond, T. (2017). Investigating young children's human figure drawings using Rasch analysis. *Educational Psychology, 37*(7), 1-19. doi:10.1080/01443410.2017.1287882

Chalmers, R. P. (2012). MIRT: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1-29. Retrieved from http://www.jstatsoft.org/v48/i06/

Chiu, T.-W., & Camilli, G. (2013). Comment on 3PL IRT adjustment for guessing. *Applied Psychological Measurement, 37*(1), 76-86. doi:10.1177/0146621612459369

Cronin, A., Gross, J., & Hayne, H. (2017). The effect of instruction on children's human figure drawing (HFD) tests: Implications for measurement. *Psychology of Aesthetics, Creativity, and the Arts, 11*(2), 179-186. doi:10.1037/aca0000097

Dans-Lopez, G., & Tarroja, M. C. H. (2010). Exploring human figure drawings as an assessment tool for departing of domestic helpers and caregivers. *Philippine Journal of Counseling Psychology, 12*(1), 13-38. Retrieved from https://www.researchgate.net/publication/228630118_Exploring_Human_Figure_Drawings_as_an_Assessment_Tool_for_Departing_OFW_Domestic_Helpers_and_Caregivers

De Ayala, R. J. (2009). *The theory and practice of item response theory.* New York: Guilford Publications.

Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for Psychologists.* New Jersey: Lawrence Erlbaum.

Flores-Mendoza, C. E., Abad, F. J., & Lelé, A.J. (2005). Análise de itens do desenho da figura humana: aplicação de TRI [Item analysis of human figure drawings test: aplication of IRT]. *Psicologia: Teoria e Pesquisa, 21*(2), 243-254. doi:10.1590/s0102-37722005000200015

Goodenough, F. L. (1926). *Measurement of intelligence by drawings.* Chicago: World Book Company

Hammer, E. F. (1958). *The clinical application of projective drawings.* Springfiled: C.C. Thomas.

Harris, D. B. (1965). Children's drawings as measures of intellectual maturity. *Journal of Aesthetics and Art Criticism, 23*(4), 516-516. doi: 10.2307/1319660

Hutz, C. & Antoniazzi, A. (1995). O desenvolvimento do desenho da figura humana em crianças de 5 a 15 anos de idade: normas para avaliação [The development of Human Figure Drawing in children of 5 to 15 years old: standards for assessment]. *Psicologia: Reflexão e Crítica, 8*(1), 3-18.

Imuta, K., Scarf, D., Pharo, H., & Hayne, H. (2013). Drawing a close to the use of human figure drawings as a projective measure of intelligence. *PLoS ONE, 8*(3), e58991. doi:10.1371/journal.pone.0058991

Koppitz, E. M. (1984). *Psychological evaluation of human figure drawings by middle school pupils.* Michigan: Grune & Stratton.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* New York: Routledge.

Machover, K. (1949). *Personality projection in the drawing of the human figure: a method of personality investigation.* Illinois: Springfield.

Magis, D. (2013). A note on the item information function of the four-parameter logistic model. *Applied Psychological Measurement, 37*(4), 304-315. doi:10.1177/0146621613475471

Muthén, B. O. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement, 28*(4), 338-354. Retrieved from http://www.jstor.org/stable/1434897

Naglieri, J. A. (1988). *DAP: Draw a person, a quantitative scoring system.* New York: Psychological Corporation.

Oliveira, K.S., & Wechsler, S. M. (2016). Indicadores de criatividade no desenho da figura humana [Creativity indicators in the Human Figure Drawing]. *Psicologia: Ciência e Profissão, 36*(1), 6-19. doi:10.1590/1982-3703001682014

Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of Personality Assessment, 92*(6), 544-559. doi:10.1080/00223891.2010.496477

Revelle, W. (2017). *Psych: Procedures for psychological, psychometric, and personality research.* Evanston, Illinois: Northwestern University. Retrieved from https://CRAN.R-project.org/package=psych

Reynolds, C. R., & Hickman, J. A. (2004). *Draw-a-person intellectual ability test for children, adolescents and adults examiner manual.* Austin, TX: Pro-ed.

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods, 21*(2), 137-150. doi:10.1037/met0000045

Silva, R.B.F., Pasa, A., Castoldi, D.R., & Spessatto, F. (2017). O desenho da figura humana e seu uso na avaliação psicológica [The Drawing of the Human Figure and its use in psychological assessment]. *Psicologia Argumento, 28*(60), 55-64. Retrieved from https://periodicos.pucpr.br/index.php/psicologiaargumento/article/download/19837/19143

Sisto, F. F. (2005). Um estudo sobre a dimensionalidade do teste do desenho da figura humana [Study about the dimensionality of the Human Figure Drawing test]. *Interação em Psicologia, 9*(1), 11-19. doi:10.5380/psi.v9i1.3282

Suehiro, A. C. B., Benfica, T. de S., & Cardim, N. A. (2016). Produção científica sobre o teste desenho da figura humana entre 2002 e 2012 [Scientific production about Teste Desenho da Figura Humana between 2002 and 2012]. *Psicologia: Ciência e Profissão, 36*(2), 439-448. doi: 10.1590/1982-3703000822014.

Wechsler, S.M. (1998). Adaptação e validação do desenho da figura humana para crianças brasileiras. *Revista Iberoamericana de Diagnóstico y Evaluación Psicológica, 4*, 47-64.

Wechsler, S.M. (2003). *DFH III: O desenho da figura humana: Avaliação do desenvolvimento cognitivo de crianças brasileiras.* Campinas, SP: Editora Da Pontifícia Universidade Católica de Campinas.

Wechsler, S. M., Prado, C.M., Oliveira, K.S., & Mazzarino, B. G. (2011). Desenho da figura humana: Análise da prevalência de indicadores para avaliação emocional. *Psicologia: Reflexão e Crítica, 24*(3), 411-418. doi: 10.1590/S0102-79722011000300001.

Wechsler, S. M., & Schelini, P. W. (2002). Validade do desenho da figura humana para avaliação cognitiva infantil [Validity of Human Figure Drawing for children's cognitive assessment]. *Avaliação Psicológica, 1*(1), 29–38. Retrieved from http://pepsic.bvsalud.org/pdf/avp/v1n1/v1n1a04.pdf

Wickham, H. (2017). *tidyverse: Easily Install and Load the 'tidyverse'.* R package version 1.2.1. https://CRAN.R-project.org/package=tidyverse

Zielona-Jenek, M. (2013). Human figure drawings in the diagnosis of child sexual offenders. potential and limitations of method in psychological forensic evaluation. *Problems of Forensic Sciences, 93*, 438-449. Retrieved from https://www.researchgate.net/publication/258112824_Human_figure_drawings_in_the_diagnosis_of_child_sexual_offenders_Potential_and_limitations_of_method_in_psychological_forensic_evaluation

recebido em setembro de 2017
aceito em setembro de 2018

Sobre os autores

Ricardo Primi é psicólogo, mestre e doutor em Psicologia pela USP-SP, professor do curso de pós-graduação *Stricto Sensu* em Psicologia da Universidade São Francisco. Bolsista Produtividade do CNPq do qual recebeu financiamento para essa pesquisa.

Tatiana de Cassia Nakano é psicóloga, mestre e doutora em Psicologia, com pós-doutorado pela Universidade São Francisco; professora do curso de pós-graduação *Stricto Sensu* em Psicologia da Pontifícia Universidade Católica de Campinas. Bolsista Produtividade do CNPq.

Solange Muglia Wechsler é psicóloga, mestre e doutora em Psicologia, com pós-doutorado pela University of Georgia (EUA) e University of Buffalo (EUA); professora do curso de pós-graduação *Stricto Sensu* em Psicologia da Pontifícia Universidade Católica de Campinas. Bolsista Produtividade do CNPq.