

UC Irvine

UC Irvine Previously Published Works

Title

Is It Worthy to Take Account of the “Guessing” in the Performance of the Raven Test? Calling for the Principle of Parsimony for Test Validation

Permalink

<https://escholarship.org/uc/item/2q70q3f4>

Journal

Journal of Psychoeducational Assessment, 39(1)

ISSN

0734-2829

Authors

Lúcio, Patrícia Silva
Vandekerckhove, Joachim
Polanczyk, Guilherme V
[et al.](#)

Publication Date


2021-02-01


DOI

10.1177/0734282920930923

Peer reviewed

Is It Worthy to Take Account of the “Guessing” in the Performance of the Raven Test? Calling for the Principle of Parsimony for Test Validation

Journal of Psychoeducational Assessment
2020, Vol. 0(0) 1–12
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0734282920930923
journals.sagepub.com/home/jpa


Patrícia Silva Lúcio¹ , Joachim Vandekerckhove²,
Guilherme V. Polanczyk³, and Hugo Cogo-Moreira^{4,5}

Abstract

The present study compares the fit of two- and three-parameter logistic (2PL and 3PL) models of item response theory in the performance of preschool children on the Raven's Colored Progressive Matrices. The test of Raven is widely used for evaluating nonverbal intelligence of factor *g*. Studies comparing models with real data are scarce on the literature and this is the first to compare models of two and three parameters for the test of Raven, evaluating the informational gain of considering *guessing* probability. Participants were 582 Brazilian's preschool children ($M_{\text{age}} = 57$ months; $SD = 7$ months; 46% female) who responded individually to the instrument. The model fit indices suggested that the 2PL fit better to the data. The difficulty and ability parameters were similar between the models, with almost perfect correlations. Differences were observed in terms of discrimination and test information. The principle of parsimony must be called for comparing models.

Keywords

Raven, model selection, item response theory, Akaike information criterion, Bayesian information criterion

Introduction

Item response theory (IRT) belongs to a class of model-based measurement, which estimates trait level of some construct based on the pattern of person's response to items and the items' properties itself (Lord, 1980). Such models allow estimating the probability of endorsing the item given the person's trait level (Reise et al., 2005). Due to its advantages over the classical test

¹State University of Londrina, Brazil

²University of California, Irvine, USA

³University of Sao Paulo, Brazil

⁴Federal University of Sao Paulo, Brazil

⁵Freie Universität Berlin, Germany

Corresponding Author:

Patrícia Silva Lúcio, Departamento de Psicologia e Psicanálise, Universidade Estadual de Londrina, Avenida Celso Garcia Cid, PR445, Campus Universitário, Londrina, CEP 86.057-970, Brazil.

Email: pslucio@gmail.com

theory, in the last years IRT became the mainstream for measurement tests and expanded its influence to other areas beyond cognitive and personality assessment, such as organizational and clinical settings (e.g., Foster et al., 2017; Reise & Waller, 2009).

Traditional logistic models family are composed by the one-parameter logistic model (1PL; also known as Rasch model), the two-parameter logistic (2PL) model, and for the three-parameter logistic (3PL) model (Lord, 1980). Roughly, the 1PL estimates the probability of endorsing an item based on its difficulty, b , as compared with each person's trait level (i.e., holding the amount of trait level needed for passing or endorsing the item). In this model, discrimination (a) is equal among the items, differently from the 2PL model, in which discrimination is freely estimated. According to Reise et al. (2005, p. 95), "more discriminating items are better able to differentiate among individuals in the trait range around an item's difficulty" (i.e., at b). Finally, the 3PL adds the parameter c or guessing. It is appropriate for multiple-choice tests, in which the probability of success from a very low-ability person in an item may be significantly higher than zero because of random guessing (Diamond & Evans, 1973) or other factors such as plausibility of distractors (De Mars, 2010).

Two- and three-parameter models have clear advantages in relation to the 1PL, because 1PL models present a strong assumption that the items present the same discrimination (Traub, 1983). Both 2PL and 3PL models are suitable for cognitive tasks such as the Raven's intelligence test (Raven et al., 2003). In this instrument, it is supposed that a single factor, g , underlies the subject's performance on the task (i.e., calling upon the unidimensionality assumption), which in turn is composed by a set of multiple choice nonverbal problems which are dichotomously scored as correct or incorrect. Comparing competing models via their model fit indices is important for building up coherent assumptions for the reality, mainly for psychology, which is primarily concerned to discovering plausible explanations about human behavior (Vandekerckhove et al., 2015). Therefore, an accurate interpretation of the data depends closely on the choice of the model that will represent it.

We found out only one study comparing the applicability of different IRT models for the Raven's test. Van der Elst et al. (2013) investigated the psychometric properties of the shortened version of the Raven (Standard Progressive Matrices) in a sample of health adults. The authors compared the responses under the 1PL and 2PL models and demonstrated that the estimated IQ was very similar under both methods (an almost perfect correlation of .97). Nevertheless, the 2PL produced better reliability indicators than the 1PL, especially considering IQ estimate range between 75 and 110, what was attested by analysis of test information function curves and of estimated reliabilities $r = .90$ and $.80$ for 2PL and 1PL, respectively.

Beyond the work of Van der Elst et al. (2013), we were able to find out only one study with real data whose intention was to compare fit index under different IRT models. Chernyshenko et al. (2001) compared the fit index of two well-known personality tests, the Sixteen Personality Factor Questionnaire (16PF) and the Big Five Personality Factor Scales under various IRT models including 2PL and 3PL for dichotomously scored items. The results indicated that both 2PL and 3PL models produced inconsistent χ^2/df fit index (i.e., the goodness of fit depended on the subscale considered). The authors discussed about the relevance of comparing models for attesting the validity of instruments that assess psychological constructs, especially in noncognitive context.

The present study aims to compare the 2PL and 3PL under the one-dimensional (1D) models of IRT in the performance of a sample of preschool children on the Raven's Colored Progressive Matrices (CPM). A previous study (Lúcio et al., 2019) showed that the intended theoretical structure of the instrument (i.e., the general structure of the nonverbal intelligence or g -factor) fit to the data in a 2PL 1D model. Nevertheless, as the test of Raven comprises multiple-choice items, the process of choosing the correct answer is susceptible to guessing (what can overestimate the abilities of low ability subjects and those which perform random response). Therefore, it is relevant to verify how worthwhile is sacrificing parsimony (i.e., choosing a model more complex

or with more parameters) in favor of the best-fit model (i.e., if the 3PL prove to be the best model), even if the data produced are not contradictory in relevant ways (Vandekerckhove et al., 2015).

Using real data for comparing competing IRT models are important for some reasons. First, the estimates of item parameter and person abilities based on IRT mathematical models are underpinned on a testable theory (Reise, 2015). In other words, the assumption that an unobserved latent variable (or ability) explains the pattern of performance in a set of items may be formally tested with the comparison of alternative models that could explain this pattern. Therefore, if different models may represent the construct, it is important to formally test what models better fit the data. This statement leads us to a second reason: as Hambleton (1994) remind us, models are not correct or wrong, but they are or not useful for representing data. In other words, models are used to explain or fit the data; therefore, scientists should pursue the best model among an outspread of possibilities. Finally, and specifically talking about comparison between 2PL and 3PL models (the matter of the present article), although both models present the same functional structure (i.e., 1D), the models differ in terms of complexity, calling upon parsimony issues when evaluating the best model (Bonifay & Cai, 2017; Vandekerckhove et al., 2015). As Bonifay and Cai (2017) discuss, complexity should not be judged based uniquely on the number of free parameters, so other aspects of the models beyond the fit index should be taken in consideration for model comparison. Therefore, both qualitative and quantitative elements should be considered when comparing models (Vandekerckhove et al., 2015).

Method

Sample

The sample of this study comes from the baseline measures of the CPM from a cluster randomized clinical trial (NCT02807831) designed to evaluate the effects of two interventions among preschoolers: the oral language and the executive functions interventions compared with a control group. The 582 preschoolers that composed the sample were randomly picked from 27 schools, nine from each group. Age varied from 43 months to 73 months ($M_{\text{age}} = 57$ months and $SD = 7$ months; 46% female).

Instrument

The Brazilian version of the CPM was used (Angelini et al., 1999; Paula et al., 2018). The Raven's CPM (Raven et al., 2003) is a nonverbal intelligence test composed by 36 items, distributed in three sets of 12 items (series A, Ab, and B). For each item, the subject must choose the missing part that completes the pattern of one picture, being one correct response in six options. Although there are six options of responses, the items are scored dichotomously (i.e., correct responses are scored with 1 point and wrong responses with 0 point), so that the maximum total score in the original version of the instrument is 36. A previous study (Lúcio et al., 2019) showed that for the sample of the present study (i.e., preschool children), the 1D model fit better for the six first items of each series of the instrument and thus the other items were removed for score composition (what produced a total score of 18). Therefore, in this study we used this score composition for the models comparisons (2PL and 3PL under 1D models).

Procedures

This study adheres to the ethical standards for research involving human being, as recommended by the Committee on Publication Ethics (COPE). Only the children whose parents provided written consent participated in the study. Trained psychologists tested the children individually in a quiet room of their schools, according to instructions given in the test manual.

Statistical Analysis

The IRT analysis was performed using the *Mplus* statistical program version 8.1 (Muthén & Muthén, 2018) and classical analysis with SPSS 25.0. For the IRT, based on monotone homogeneity (see Mokken, 1971), where the item characteristic curves can differ from an item to the next, the following assumptions were considered: local independence was checked via bivariate Pearson standardized residuals, z -score (Agresti, 2019; Haberman, 1973). Traditionally, z -scores exceeding $|1.96|$ would indicate violations of local independence (i.e., reject the null hypothesis of local independence). We computed standardized Pearson residuals (Haberman, 1973), which are normally distributed z -scores.

Because the children were nested within 27 schools, the models take into account such nonindependence where standard errors and a chi-square test of model fit were computed considering such multilevel structure by command in *Mplus* called (TYPE = Complex) as proposed by Asparouhov (2005, 2006); standard error computations use a sandwich estimator. The 2PL and the 3PL under 1D models were performed using robust maximum likelihood estimator. *Mplus* implemented a prior maximum likelihood parameter that helps 3PL model convergence (see Asparouhov & Muthén, 2016), whereas previously large sample size (>1,000 participants) was necessary to run 3PL. Because the Raven has six options of responses, the a priori probability (i.e., prior) for the guessing parameter was $1/6$ or .1667.

For model fit, we used the Akaike and Bayesian information criteria (respectively, AIC and BIC). Under maximum likelihood, both 2LP and 3PL do not generate the traditional model fit index such as chi-square, comparative fit index (CFI), Tucker–Lewis index (TLI), and weighted root mean square residual (WRMR), which are available under weighted least square mean and variance adjusted (WLSMV) estimators (the latter is the default estimator when items are categorical in *Mplus*). Nevertheless, as indicated by Sen and Bradshaw (2017), both AIC and BIC correctly select the correct model under IRT modeling. However, AIC and BIC penalize more complex models (i.e., as the number of parameters increases, some loss in goodness of fit is observed in the model). This is because both AIC and BIC indices consider the number of parameters k in their numerators (i.e., $2k$ for AIC and $k \cdot \ln^* n$ for BIC), and models with smaller AIC and BIC are selected, so it's recommended to perform adjustments. Vandekerckhove et al. (2015) presents the Akaike and Schwarz weights (respectively, wAIC and wBIC) in equation (14.4) of their chapter. This equation considers the difference between the information criteria (IC) of the models (Δ_i), that is, it considers the relative performance and not the absolute AIC or BIC values. When the difference between the two model AIC or BIC scores is more than 20, the wAIC and wBIC will be near 1 for the better model and 0 for the others. (Because we are comparing only two models, we performed the difference between the AIC and BIC of the 2PL and 3PL models. When there are more models to be compared, we should consider the difference between the IC of all models with the IC of better model, that is, the one that has the lower IC. For more details, see Vandekerckhove et al., 2015.)

Information test curves were exposed, and for ability estimate (latent trait or person-fit) *Mplus* SAVE = *f*scores function were used. Descriptive statistics were displayed for discrimination, difficulty, ability, and raw scores. Pearson correlations and t test for related samples were performed for comparisons between items parameters.

Results

Figure 1 shows the distribution of the standardized residual (z -scores), with 612 bivariate residuals. As can be inspected by the picture, most of the bivariate standardized residuals are around zero, ranging from -2.10 to 2.00 ($M = -.01$; $SD = .659$). Therefore, no meaningful deviations were observed, meaning no evidence for local independences violations.

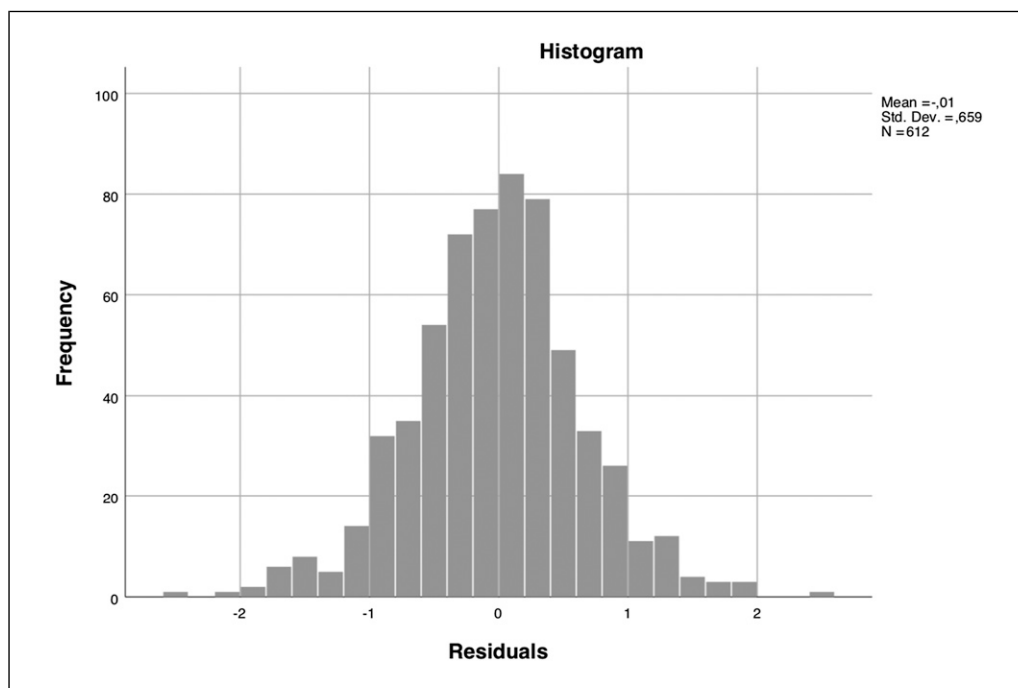


Figure 1. Distribution of the Pearson standardized residuals (z-score).

For the 1D model of the 18 items of the Raven test (see Lúcio et al., 2019, Figure 1, for the model representation), the 2PL presented better fit than the 3PL based on AIC and BIC (2PL—AIC: 10,723.884; BIC: 10,881.077 and 3PL—AIC: 10,816.840; BIC: 11,052.629). The differences between the scores were far away from 20 for both indices (respectively, 92.956 for AIC difference and 171.549 for BIC difference, both greater for 3PL). Using equation (14.4) of the chapter of Vandekerckhove et al. (2015), the wAIC for the 2PL was $1 - 2E-19$ and the wAIC for the 3PL was approximately $2E-19$. The difference in the BIC scores was much bigger still, so the wBIC for 2PL was essentially 1 and $3.6E-218$ for the 3PL. Therefore, using these weighted indices, the 2PL was the better model.

Figure 2(A) and (B) depicts the information test curve for the 2PL and the 3PL models, respectively. As can be observed in Figure 2(A), for the 2PL the test is more informative for the g -factor level between -2.0 and -0.5 . Otherwise, Figure 2(B) presents a bimodal-like curve with a peak of g -factor level between -2.0 and -0.5 and another (less informative) peak between 0.5 and 2.5 .

Table 1 presents IRT 1D index of discrimination and difficulty of items for the 2PL and 3PL models and the guessing parameter for the 3PL, as their respective standard errors. Bivariate correlations (Pearson) showed significant ($p < .001$) relationship between discrimination index ($r = .75$) and difficulty index ($r = .99$) of the two models (2PL and 3PL). The t tests for paired samples showed significant differences between mean discrimination, $t(17) = 2.549$, $p = .021$, $d = 0.42$, with higher discrimination for 3PL, and no differences between difficulty means, $t(17) = 1.579$, $p = .133$, $d = 0.05$.

Descriptive statistics for ability estimate for the 2PL and 3PL models and for raw scores are presented in Table 2. The Pearson correlations between the ability estimates achieved for both models were high ($r = .996$, $p < .001$), and this correlation is depicted in Figure 3. The t -test difference between the person fit obtained by the two procedures was not significant,

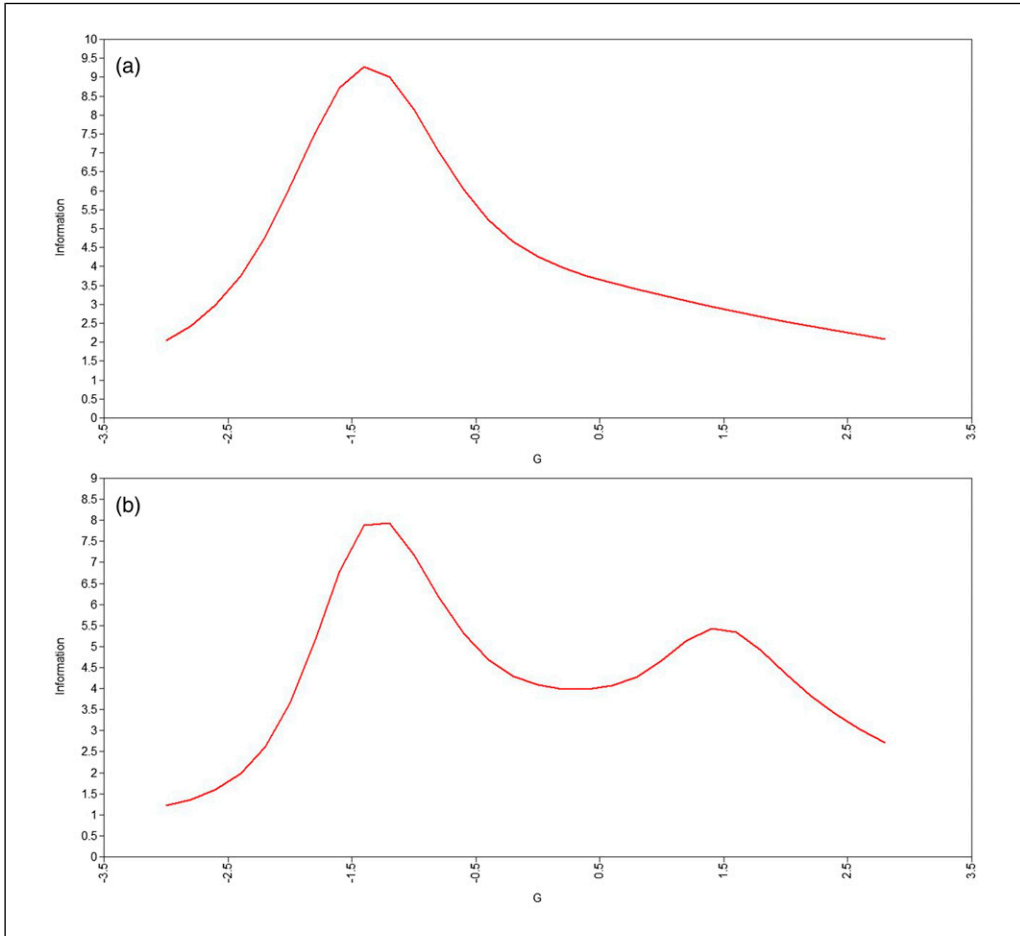


Figure 2. Information test curve for the (a) 2PL and the (b) 3PL models.
 Note. 2PL = two-parameter logistic; 3PL = three-parameter logistic.

$t(581) = 0.151, p = .880, d < 0.001$. As we can see in [Figure 4\(A\) and \(B\)](#), the latent trait distribution obtained is very similar through both 2PL and 3PL. The correlations between the raw score and the ability obtained through the 2PL ($r = .986, p < .001$) and the 3PL ($r = .978, p < .001$) were also high, as expected.

Discussion

The present study aimed to compare the application of two equally structured models (i.e., 1D) with different number of parameters, namely, the 2PL and the 3PL models of IRT, to a traditional instrument of assessment of nonverbal ability. We used the reduced form of the Raven's CPM suitable for a sample of preschoolers ([Lúcio et al., 2019](#)), composed by the first six items of each of the three series of the test (totalizing 18 items). Other studies confirmed that the reduced forms of the Raven's matrices produce results with similar psychometric properties to the long forms (e.g., [Arthur & Day, 1994](#); [Bilker et al., 2012](#)). Using maximum likelihood estimators we performed difficulty, discrimination, and ability estimates for both models and guessing parameter for the 3PL model.

Table 1. IRT Parameters for the 2PL and 3PL Models Obtained for the Reduced Form of the CPM.

Item	2PLM				3PLM					
	<i>a</i>	<i>SE</i>	<i>b</i>	<i>SE</i>	<i>a</i>	<i>SE</i>	<i>b</i>	<i>SE</i>	<i>c</i>	<i>SE</i>
A1	1.094	0.138	-0.203	0.166	1.294	0.212	0.069	0.218	0.119	0.045
A2	3.183	0.678	-1.523	0.159	3.621	0.931	-1.444	0.151	0.102	0.021
A3	2.025	0.328	-1.678	0.148	1.990	0.341	-1.637	0.145	0.095	0.012
A4	1.981	0.260	-1.400	0.138	1.968	0.259	-1.356	0.127	0.078	0.012
A5	0.895	0.157	-0.474	0.168	0.901	0.160	-0.323	0.158	0.068	0.012
A6	1.059	0.148	0.165	0.122	1.250	0.206	0.383	0.135	0.095	0.023
AB1	1.610	0.222	-0.987	0.136	1.633	0.252	-0.890	0.140	0.083	0.015
AB2	0.931	0.154	-0.081	0.132	1.049	0.177	0.186	0.135	0.104	0.028
AB3	0.988	0.169	0.031	0.145	1.089	0.195	0.249	0.167	0.088	0.022
AB4	0.734	0.129	1.927	0.347	0.985	0.205	1.983	0.330	0.073	0.020
AB5	0.827	0.130	0.485	0.144	0.924	0.151	0.695	0.154	0.076	0.015
AB6	0.797	0.156	2.346	0.415	1.341	0.339	2.116	0.300	0.070	0.019
B1	2.931	0.564	-1.205	0.157	2.891	0.576	-1.180	0.151	0.073	0.018
B2	0.734	0.168	1.432	0.333	3.131	2.857	1.402	0.165	0.194	0.052
B3	0.790	0.141	0.422	0.174	1.085	0.204	0.797	0.251	0.144	0.059
B4	1.082	0.179	0.257	0.185	1.491	0.432	0.536	0.226	0.130	0.076
B5	0.777	0.176	1.985	0.389	1.564	0.620	1.786	0.260	0.100	0.037
B6	0.630	0.181	2.589	0.622	1.255	0.392	2.164	0.382	0.090	0.033
Mean	1.282	0.227	0.227	0.227	1.637	0.473	0.308	2.000	0.099	0.029
SD	0.764	0.153	1.136	0.137	0.802	0.628	1.262	0.076	0.032	0.018

Note. *a* = discrimination; *b* = difficulty; *c* = guessing; IRT = item response theory; 2PL = two-parameter logistic; 3PL = three-parameter logistic; CPM = colored progressive matrices; PLM = parameter logistic model; *SE* = standard error.

Table 2. Descriptive Statistics for IRT Person Fit (Ability and Standard Error) for 2PL and 3PL Models and for Raw Scores Derived From the Reduced Form of the CPM.

Variable	Minimum	Maximum	Mean	SD
2PL score	-2.400	2.440	-0.001	0.878
2PL SE	0.350	0.660	0.477	0.066
3PL score	-2.270	2.510	-0.001	0.885
3PL SE	0.370	0.610	0.470	0.040
Raw score	0.000	18.00	9.225	3.423

Note. IRT = item response theory; 2PL = two-parameter logistic; 3PL = three-parameter logistic; CPM = colored progressive matrices; *SD* = standard deviation; *SE* = standard error.

The AIC and BIC fit index showed that the 2PL better fit the data than the 3PL. This result is not a surprise, given the tendency of both indices in penalizing model complexity because of the presence of the number of adjustable parameters as multipliers in their respective formulas (Vandekerckhove et al., 2015). Moreover, these crude indices do not access functional form issues of the models, which brings concerns about considering complexity in terms of counting parameters (see Bonifay & Cai, 2017, for a discussion). Therefore, in the present study we used other procedures to compare the models. The first step was calculating the wAIC and wBIC scores (i.e., a weighting measure based on the AIC and BIC scores of both compared models). Both wAIC and wBIC were almost 1.0 for 2PL and essentially 0.0 for the 3PL, indicating that the 2PL presented the best model-fit for this criterion.

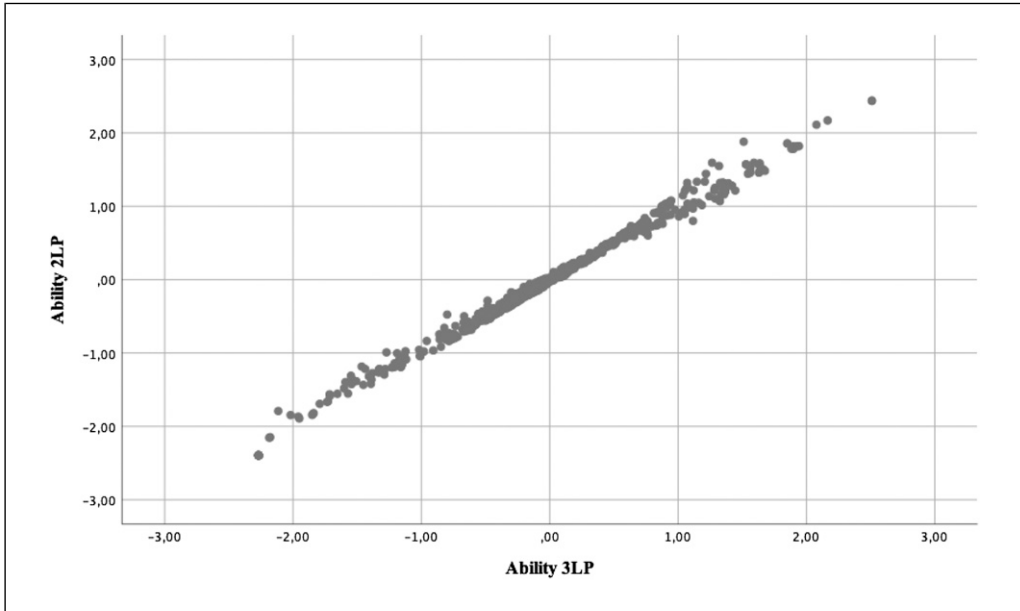


Figure 3. Correlation between latent traits (ability) obtained through the 2PL and the 3PL models.
 Note. 2PL = two-parameter logistic; 3PL = three-parameter logistic.

The graphic analysis of the test information curves (Figure 2(A) and (B)) showed that both models are more informative for the individuals with nonverbal ability lower than the mean, that is, the information peaked at latent trait between -2.5 and -0.5 (with information level around 9.5 for 2PL and 8.5 for 3PL). This is the range where the measurement error is lower for the instrument (Embretson & Reise, 2000). The greatest difference between the models was that the 3PL presented an additional peak of information (around 5.5), at latent trait between 0.5 and 2.5, meaning that this model covered a wider range of ability levels with measurement precision. We were not able to find out similar analysis for the Raven's CPM, but using a sample of Italian and English adults, Chiesi et al. (2012) showed that the information function of the short form of the Advanced Progressive Matrices peaked at latent trait between -1.0 and 1.0 in a three-parameter model. In the Van der Elst et al. (2013) study with Dutch adults the test information of 2PL was higher for ability estimates between -2.0 and 0.0 , using the short version of the Standard Progressive Matrices. Despite the differences between samples and tests, it seems that the test information for the 2PL may be higher for lower ability individuals whereas the test information for the 3PL reaches peak at a wider range of abilities. More studies are necessary for tracing a picture of modeling differences in test information.

As expected, the discrimination index were significantly different between the two models (higher for the 3PL), despite the correlation was pretty high (.75). As the 3PL adds a probability of correcting response by chance, the lower asymptote a prior probability of not being 0.0 at the Y-axis and, therefore, some differences are expected in discrimination. The addition of the guessing parameter may have increased the item's capacity of distinguishing between person's ability of correctly endorsing items (i.e., a). For Table 2, it is possible that these differences are loaded at the higher difficult items (series B), which are more susceptible to guessing. Thus, the capacity of estimating guessing may have produced higher discrimination index for the 3PL. Contrariwise, no differences between the models were observed between the difficulty indices and the persons' fit. This suggests that both models are accurately well in determining the amount of

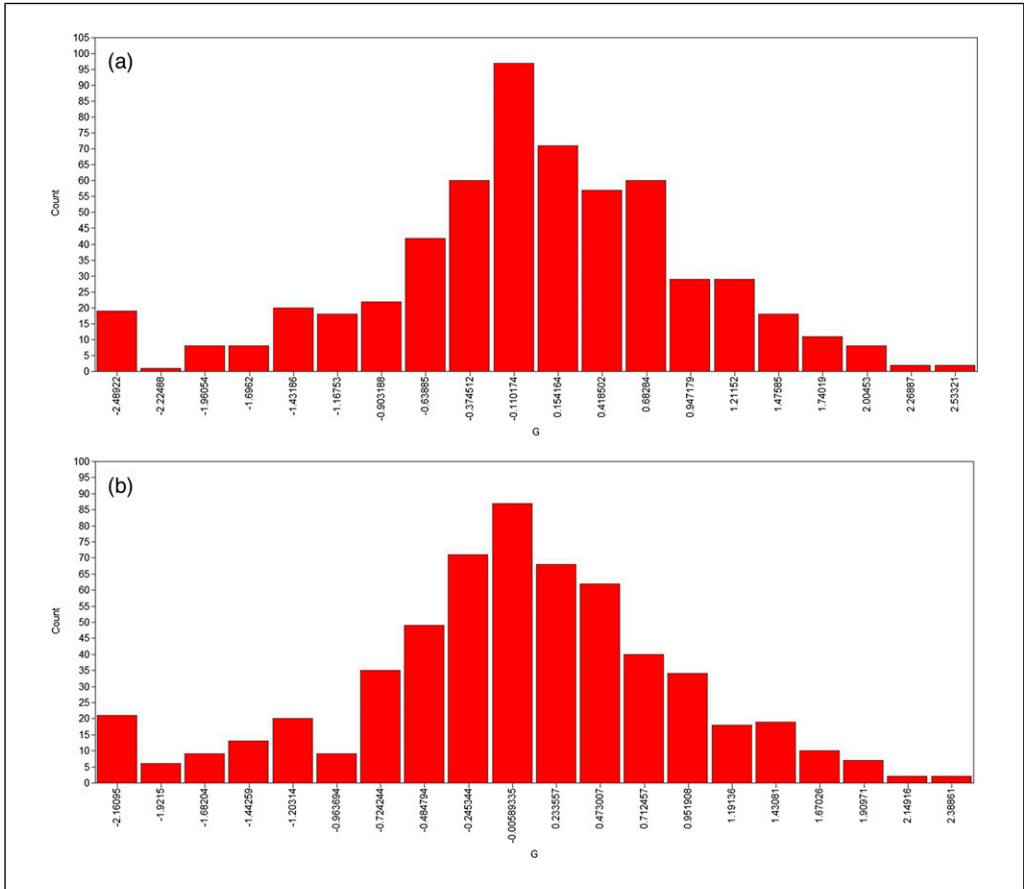


Figure 4. Latent traits distribution for the sample with the (a) 2PL and (b) 3PL models (note that the count in the Y-axis is slightly different in both (a) and (b)).
 Note. 2PL = two-parameter logistic; 3PL = three-parameter logistic.

latent trait necessary to endorse the item (i.e., *b*) and the estimate of persons’ ability based on the response pattern (i.e., latent trait).

The approach of estimating the *c* or guessing parameter seemed to be adequate for the 3PL ($M = 0.099$; $SD = 0.029$). As De Mars (2010) point out, the guessing parameter can fall below or above that $1/\text{number of options of the test}$ (i.e., not always guessing presents a literal random meaning, because different factors may improve or decrease the chance of endorsing an item entirely by random). Typically, *c* is lower than $1/\text{number of options}$ in well-developed tests, probably because distractors functions correctly and people with low-ability may choose the correct response fewer times than by chance. For the Raven test, the expected probability of answer by chance would be .1667 (because there are six options of answers). In the present research, the freely estimated guessing (*c*) parameter was not higher than .14 (item B3).

Finally, we should point some limitations and future developments for this research. First, we contrasted 2PL and 3PL models because they are readily related to the scores derived from the Raven test: it presents a 1D construct, its items are dichotomously corrected, and it is susceptible to guessing (because it is multiple-choice test). Another possibility for data analysis is, for example, using IRT for polytomous items such as the graded response model (Samejima, 1968). In

the Raven test, there are categories of responses that could be associated to a certain pattern of reasoning (e.g., the respondent can mismatch the target because he or she presents difficulty in encoding relevant information for the problem-solving). This is an application that could be performed in a more integrative view of the responses to the test (and therefore for the abilities of the respondents). Second, some observations should be done about the use of IRT-theta scale as score scale, which may limit the extent of our results. Kolen and Brennan (2014) question the utility of such scale for paper and pencil tests because ability derivation depends on the pattern of answers, what can be derived for computerized tests (i.e., two individual with the same total scores may present different latent level because of the kind of items they corrected or missed). However, the raw score is highly correlated with latent ability, as demonstrated in this study for the almost perfect correlation for both models (.986 for the 2PL and .978 for the 3PL). Therefore, raw score may be a proxy of the ability and, thus, there is some utility in deriving this measure for the models under interest. Finally, another issue put by Kolen and Brennan is trickier. The authors point out that under IRT the measurement error is typically greater for examinees with extreme scores, which makes latent trait estimates less reliable for these groups than classic score-derivatives, such as z score. The stability of latent trait estimates for extreme groups, and hence our reliance on the interpretation of summary statistics and the correlations derived from these data, should be explored in future research.

Conclusion

The present study compared the 2PL and 3PL 1D logistic models of IRT in the performance of preschoolers from Brazil on the Raven's CPM. The instrument is suitable for both models because it supposes that a single latent trait (i.e., g factor) underpins the performance and is composed by dichotomously corrected multiple-choice items. Although the AIC and BIC estimators showed the superiority of 2PL under the 3PL, as its weighted derived measures (i.e., wAIC and wBIC), the comparison of the parameters generated showed that the models were quite similar, with some superiority of the 3PL in regard of the range of coverage of the test information function and the discrimination parameter. We conclude that both models fit well, but for the principle of parsimony it's worthwhile using the 2PL. For the practitioners, this found is relevant because it adds information about the construct validity of the instrument, meaning that the interpretations derived from the items' scores are in some extent trustworthy. In other words, guessing may be not relevant for score composition. This is a well-come result for users, mainly considering that the peak of information of the instrument (i.e., greater reliability) is at lower trait levels, were guessing is more susceptible to take place.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by Fundação de Amparo à Pesquisa de São Paulo (FAPESP; grant number 12/51624-1).

ORCID iD

Patrícia Silva Lúcio  <https://orcid.org/0000-0001-7125-206X>

References

- Agresti, A. (2019). *An introduction to categorical data analysis* (3rd ed.). John Wiley & Sons.
- Angelini, A. L., Alves, I. C. B., Custódio, E. M., Duarte, W. F., & Duarte, J. L. M. (1999). *Raven's coloured progressive matrices: Special scale*. Centro Editor de Testes e Pesquisas em Psicologia.
- Arthur, W., Jr., & Day, D. V. (1994). Development of a short form for the Raven advanced progressive matrices test. *Educational and Psychological Measurement*, *54*(2), 394-403.
- Asparouhov, T. (2005). Sampling weights in latent variable modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *12*(3), 411-434. doi:10.1207/s15328007sem1203_4
- Asparouhov, T. (2006). General multi-level modeling with sampling weights. *Communications in Statistics—Theory and Methods*, *35*(3), 439-460. doi:10.1080/03610920500476598
- Asparouhov, T., & Muthén, B. (2016). *IRT in Mplus*. <https://www.statmodel.com/download/MplusIRT.pdf>
- Bilker, W. B., Hansen, J. A., Brensinger, C. M., Richard, J., Gur, R. E., & Gur, R. C. (2012). Development of abbreviated nine-item forms of the Raven's standard progressive matrices test. *Assessment*, *19*(3), 354-369.
- Bonifay, W., & Cai, L. (2017). On the complexity of item response theory models. *Multivariate Behavioral Research*, *52*(4), 465-484. doi:10.1080/00273171.2017.1309262
- Chernyshenko, O. S., Stark, S., Chan, K. Y., Drasgow, F., & Williams, B. (2001). Fitting item response theory models to two personality inventories: Issues and insights. *Multivariate Behavioral Research*, *36*(4), 523-562. doi:10.1207/S15327906MBR3604_03
- Chiesi, F., Ciancaleoni, M., Galli, S., Morsanyi, K., & Primi, C. (2012). Item response theory analysis and differential item functioning across age, gender and country of a short form of the advanced progressive matrices. *Learning and Individual Differences*, *22*(3), 390-396. doi:10.1016/j.lindif.2011.12.007
- De Mars, C. E. (2010). Guessing parameter. In N. J. Salkind (Ed.), *Encyclopedia of research design* (pp. 557-558). SAGE.
- Diamond, J., & Evans, W. (1973). The correction for guessing. *Review of Educational Research*, *43*(2), 181-191. doi:10.3102/00346543043002181
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum.
- Foster, G. C., Min, H., & Zickar, M. J. (2017). Review of item response theory practices in organizational research: Lessons learned and paths forward. *Organizational Research Methods*, *20*(3), 465-486. doi:10.1177/1094428116689708
- Haberman, S. J. (1973). The analysis of residuals in cross-classified tables. *Biometrics*, *29*(1), 205-220. doi:10.2307/2529686
- Hambleton, R. K. (1994). Item response theory: A broad psychometric framework for measurement advances. *Psicothema*, *6*(3), 535-556.
- Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). Springer.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum.
- Lúcio, P. S., Cogo-Moreira, H., Puglisi, M., Polanczyk, G. V., & Little, T. D. (2019). Psychometric investigation of the Raven's colored progressive matrices test in a sample of preschool children. *Assessment*, *26*(7), 1399-1408.
- Mokken, R. J. (1971). *A theory and procedure of scale analysis, volume 1 of methods and models in the social sciences*. The Gruyter.
- Muthén, L. K., & Muthén, B. O. (2018). *Mplus. The comprehensive modelling program for applied researchers: User's guide* (Version 8).
- Paula, J. J., Schlottfeldt, C. G. M. F., Malloy-Diniz, L. F., & Mizuta, G. A. A. (2018). *Raven's coloured progressive matrices: Manual*. Pearson Clinical Brasil.
- Raven, J., Raven, J., & Court, J. (2003). *Manual for Raven's progressive matrices and vocabulary scales*. Oxford Psychologists Press.
- Reise, S. P. (2015). Item response theory. In R. L. Cautin & S. O. Lilienfeld (Eds.), *The encyclopedia of clinical psychology* (pp. 1-10). John Wiley & Sons. doi:10.1002/9781118625392.wbecp357
- Reise, S. P., Ainsworth, A. T., & Haviland, M. (2005). Item response theory: Fundamentals, applications, and promise in psychological research. *Current Directions in Psychological Science*, *14*(2), 95-101. doi:10.1111/j.0963-7214.2005.00342.x

- Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology, 5*, 27-48. doi:[10.1146/annurev.clinpsy.032408.153553](https://doi.org/10.1146/annurev.clinpsy.032408.153553)
- Samejima, F. (1968). Estimation of latent ability using a response pattern of graded scores. *ETS Research Bulletin Series, 1*(1), i-169. doi:[10.1002/j.2333-8504.1968.tb00153.x](https://doi.org/10.1002/j.2333-8504.1968.tb00153.x)
- Sen, S., & Bradshaw, L. (2017). Comparison of relative fit indices for diagnostic model selection. *Applied Psychological Measurement, 41*(6), 422-438. doi:[10.1177/0146621617695521](https://doi.org/10.1177/0146621617695521)
- Traub, R. E. (1983). A priori considerations in choosing an item response model. In R. K. Hambleton (Ed.), *Applications of item response theory* (pp. 57-70). Educational Research Institute of British Columbia.
- Vandekerckhove, J., Matzke, D., & Wagenmakers, E. J. (2015). Model comparison and the principle of parsimony. In J. R. Busemeyer, Z. Wang, J. T. Townsend, & A. Eidels (Eds.), *The Oxford handbook of computational and mathematical psychology* (pp. 300-319). Oxford Library of Psychology.
- Van der Elst, W., Ouwehand, C., van Rijn, P., Lee, N., Van Boxtel, M., & Jolles, J. (2013). The shortened Raven standard progressive matrices: Item response theory-based psychometric analyses and normative data. *Assessment, 20*(1), 48-59. doi:[10.1177/1073191111415999](https://doi.org/10.1177/1073191111415999)