Intelligence 68 (2018) 109–116

Contents lists available at ScienceDirect

Intelligence

journal homepage: www.elsevier.com/locate/intell

A snapshot of *g*? Binary and polytomous item-response theory investigations of the last series of the Standard Progressive Matrices (SPM-LS)

Nils Myszkowski[a,*], Martin Storme[b]

[a] Department of Psychology, Pace University, USA
[b] Laboratoire Adaptations Travail-Individu, Université Paris Descartes – Sorbonne Paris Cité, France

ARTICLE INFO

Keywords:
Item-Response Theory
General Mental Ability
Intelligence
Psychometrical investigation

ABSTRACT

Raven's progressive matrices (Raven, 1941) are extremely popular measures of general mental ability. However, their length may not suit every researcher's or practitioner's needs. Short versions of the Advanced version have resulted in problematic factor structures and internal consistencies (Arthur, Tubre, Paul, & Sanchez-Ku, 1999; Bors & Stokes, 1998): Is the last series of the Standard Progressive Matrices a more viable option? The aim of this research was to investigate the structural validity and internal reliability of the last series of the SPM (SPM-LS) as a standalone measure. The SPM-LS binary (correct/incorrect) responses of 499 undergraduate students were investigated through unidimensional Item-Response Theory (IRT) 1–4 Parameter Logistic (PL) models. They were satisfactorily modeled by unidimensional models (CFI_{3PL} = 0.974, TLI_{3PL} = 0.959, $RMSEA_{3PL}$ = 0.059, $SRMR_{3PL}$ = 0.056), offering good empirical reliability ($r_{xx',3PL}$ = 0.843), and outperforming the short Advanced Progressive Matrices' previously reported qualities. Full nominal responses – recovering information from the distractor responses – were further examined with recently introduced 2–4 Parameter Logistic Nested (PLN) models (Suh & Bolt, 2010), providing significant reliability gains ($\Delta r_{xx'}$ = 0.029, Bootstrapped 95% CI [0.019, 0.036], z = 6.06, $p < .001$). Uses, limitations, conditional reliability and scoring strategies are further examined and discussed.

1. Introduction

Raven's Standard Progressive Matrices (SPM) test (Raven, 1941) – along with the Advanced Progressive Matrices (APM) test – are certainly among the most heavily used and most easily administered measures of general mental ability (Pind, Gunnarsdóttir, & Jóhannesson, 2003). Because of their non-verbal content, and thus low culture loading, and because of their correlations with multidimensional measures of intelligence (e.g., Jensen, Saccuzzo, & Larson, 1988), they are often considered as central among other measures of cognitive ability (e.g., Carpenter, Just, & Shell, 1990). Although they have been criticized for not imperfectly reflecting pure *g* (Gignac, 2015), and although the form of analytic intelligence measured by Raven's matrices may to some extent be multifaceted (Carpenter et al., 1990), Raven's matrices are often considered as one of the purest measures of *g* or fluid intelligence.

A limitation of these tests is their length: The SPM is composed of 60 items and the APM is composed of 36 more complex items. Although conveniently administered, they may not fit well in one-hour research protocols that aim to administer other measures (Arthur, Tubre, Paul, & Sanchez-Ku, 1999). Accounting for these length issues, the APM have been primarily studied as candidates for two shortened 12-item versions (Arthur et al., 1999; Arthur & Day, 1994; Bors & Stokes, 1998), through two different methods – one selecting items systematically from all series (Arthur & Day, 1994), the other basing itself on item-total correlations (Bors & Stokes, 1998).

1.1. The shortcomings of the short APM

However, there are various limitations to these two APM versions. First, the existing psychometric investigations of these scales have produced mixed results, with especially somewhat problematic internal consistencies – with variable Cronbach's αs, ranging from 0.58 (Ablard & Mills, 1996) to 0.73 (Bors & Stokes, 1998) – and problematic fit of a unidimensional factor structure – with Comparative Fit Indices around 0.90 (Bors & Stokes, 1998), Tucker-Lewis Indices ranging from 0.84 to 0.87 (Arthur et al., 1999), and Non-Normed Fit Indices around 0.88 (Bors & Stokes, 1998). Because of the short length of these versions, their qualities have been considered acceptable, but the estimates reported are actually typically considered as insufficient (e.g., Hu & Bentler, 1999). In addition, Bors and Stokes (1998) actually report a 2-correlated factors solution to fit the data slightly better than a one-

* Corresponding author.
E-mail address: nmyszkowski@pace.edu (N. Myszkowski).

https://doi.org/10.1016/j.intell.2018.03.010
Received 17 October 2017; Received in revised form 15 March 2018; Accepted 26 March 2018

factor solution, further questioning the tests' unidimensionality. The authors trace back this issue to the original APM test itself, whose unidimensionality was also questioned (Dillon, Pohlmann, & Lohman, 1981).

Second, all of the investigations of these short measures have used Classical Test Theory (CTT) approaches – based on the matrix of correlations between items ("limited information"), not the raw responses ("full information"). CTT is especially pointed out as lacking in the study for ability measures of varying difficulty levels (Macdonald & Paunonen, 2002). More generally, CTT is often pointed as deficient compared with Item-Response Theory (IRT) in the study of psychological measures, for the reason that CTT constrains the study of the relation between the construct and its observations, by essentially fixing observations to reflect a "true score" with added random noise (Borsboom, 2006). In contrast, IRT allows to model responses as predicted by a non-linear function of the latent ability. In the case of categorical responses, IRT especially allows to model item responses with logistic functions of the ability. Finally, IRT allows to account for a variety of testing phenomena that CTT cannot account for, such as pseudo-guessing and slipping.

Finally, preferring the APM as a candidate for a short version is not the only available solution. Because of their complexity, the APM may appear more called for in samples of post-secondary students, but 1) post-secondary students are primarily heavily studied for convenience, and are not representative of the general population in intelligence, and 2) even in post-secondary student samples, the observed scores of the SPM are high but still allow to measure interindividual variability (e.g., Myszkowski, Storme, Zenasni, & Lubart, 2014), up to IQ levels of 134 (Jensen et al., 1988), corresponding to the 99th percentile of a normal IQ distribution. In other words, the SPM is of course easier than the APM, but its coverage is sufficient as a quickly administered IQ proxy in the normal range, as demonstrated in correlation studies between IQ measures and the test (Jensen et al., 1988). Thus, while the APM seems an optimal choice, the SPM can also be considered a good candidate for a short version, which could certainly be more appropriate for the general population, but still appropriate for post-secondary students.

1.2. Revising the SPM

In this study, we aimed at proposing a way to overcome the limitations of the existing short APM measures and of their investigations, by proposing an investigation of a short SPM version, which consists of the full most complex last set to be taken as a standalone. Additionally, since we pointed CTT assumptions to not be appropriate for Raven's matrices tests, we addressed the shortcomings of the previous investigations with the use of an alternative framework, IRT, which allows to model responses as they relate to examinee-specific latent ability and item-specific characteristics.

1.3. Studying binary responses with IRT

Because each incomplete matrix has only one correct answer and no answer considered partially correct, the typical scoring of the SPM is a count of correct answers. In other words, the categorical responses are dichotomized into pass/fail (0/1) binary responses. One of the most useful frameworks in analyzing the structure and score reliability of instruments that produce such responses is the IRT framework, which consists in building latent variable explanatory models of each item responses. In other words, IRT models the probability $p(\theta)$ of a specific response to an item as a function of the latent trait or ability (θ) – assumed constant for each test taker.

In binary IRT, responses are typically modeled as logistic functions of the ability. The resulting logistic models vary upon their number of (free) item-dependent parameters. One-Parameter Logistic (1PL) models, also named "Rasch" models, only allow items to vary by their degree of "difficulty" – or "location" – which is the point where the slope of the relation between θ and $p(\theta)$ is maximized. Two-Parameter Logistic (2PL) models allow items to vary not only by difficulty, but also by strength of relation between θ and $p(\theta)$, which is represented by a "slope" or "discrimination" parameter. Three-Parameter Logistic (3PL) models further allow items to vary by one of the asymptotes in the relation between θ and $p(\theta)$. Habitually, a variable lower asymptote is modeled, which then represents the probability to succeed an item even at very low θ levels – in other words the probability of a correct guess. For that reason, that third parameter is often referred to as a "pseudo-guessing" parameter. Finally, the 4-Parameter Logistic (4PL) model further allows items to vary by both the asymptotes in the relation between θ and $p(\theta)$. As the lower asymptote is typically estimated in 3PL models, the second asymptote modeled in 4PL models then corresponds to the probability of failing an item in spite of a very high θ. Consequently, this parameter is often called an "inattention" or "slipping" parameter.

Because the items of the SPM-LS present test takers with a set of possible answers that include the correct one, we hypothesized that guessing was not negligible – and thus should not be fixed to 0 – nor perfectly random – and thus should not be fixed to $1/8^{th}$. Therefore, we hypothesized that the 3PL and 4PL models, which estimate a pseudo-guessing parameter, would present a better fit of the responses than the 1PL and 2PL. We however hypothesized that inattention for a test of such short length would be minimal, and thus hypothesized that the 3PL model would actually fit the data better than the 4PL, which estimates an inattention parameter.

1.4. Recovering distractor information with polytomous IRT models

Although the outcomes of the SPM are typically treated as dichotomous (e.g., Lynn, Allik, & Irwing, 2004), there are actually 8 possible responses per item, and not only 2. Therefore, in reality, the SPM-LS does not directly result in binary responses, but in polytomous ones. Further, incorrect responses may include information about the ability being tested (Vodegel Matzen, van der Molen, & Dudink, 1994), suggesting that the modeling of raw polytomous responses of the SPM-LS items with polytomous IRT models could recover such information.

A heavily used IRT model of polytomous responses is the Nominal Response Model (NRM) introduced by Bock (1972), which estimates two parameters – the same parameters of location and slope than in the binary 2PL model – for each response category of each item. Suh and Bolt (2010) recently argued that the NRM model is appropriate for situations where the response to an item can be assumed to result from a comparison of all responses, but less appropriate for situations where the distractors are only considered as potential responses for test takers that are unable to find the correct response – in other words, for such situations, if the test taker follows the correct strategy, the correct response is found and the distractors are overlooked, while, if the taker is unable to follow the correct strategy, then and only then the examinee considers distractors are potential responses. For these situations, Suh and Bolt (2010) adapted Nested Logit Models (NLM) – which model nominal outcomes as conditional upon a series of choices – to Item-Response Theory. IRT NLMs have two levels: Level 1 distinguishes correct from incorrect responses – for such purposes, a binary 2-4PL model is used – while Level 2 models distractor responses with a nominal model, conditional upon an incorrect response (Suh & Bolt, 2010).

In other words, as explained by its original authors (Suh & Bolt, 2010), a 3-Parameter Nested Logit (3PNL) model models the probability $P(U_{ij} = 1|\theta_j)$ than examinee j chooses the correct response for an item i as a function of examinee ability θ_j and item parameters β_i (location), α_i (slope), and γ_i (lower asymptote):

$$P(U_{ij} = 1 \mid \theta_j) = \gamma_i + \frac{1 - \gamma_i}{1 + e^{-(\beta_i + \alpha_i \theta_j)}}$$

Further, for an item with m_i distractor categories (in the case of the

SPM-LS, $m_i = 7$ for all items), the probability of choosing a distractor v, which is noted $P(U_{ij} = 0, D_{ijv} = 1 \mid \theta_j)$, is modeled as the product of the probability of an incorrect response $1 - P(U_{ij} = 1 \mid \theta_j)$ and the probability of selecting distractor v conditional upon an incorrect response $P(D_{ijv} = 1 \mid U_{ij} = 0,\ \theta_j)$. Similar to a Nominal Response Model (Bock, 1972), this conditional probability part is a function of item category parameters ζ_{iv} (intercept) and λ_{iv} (slope) of the considered category v, and of the sum across all m_i distractor categories of the propensities towards each distractor category $\sum_{k=1}^{m_i} e^{\zeta_{ik}+\lambda_{ik}\theta_j}$.

$$P(U_{ij} = 0,\ D_{ijv} = 1|\theta_j) = [1 - P(U_{ij} = 1|\theta_j)]\left[\frac{e^{\zeta_{iv}+\lambda_{iv}\theta_j}}{\sum_{k=1}^{m_i} e^{\zeta_{ik}+\lambda_{ik}\theta_j}}\right]$$

In Raven's matrices, an examinee is expected to extract a logical rule based on the observation of the incomplete matrix. The successful extraction of the logical rule then allows the examinee to retrieve the correct missing part (Carpenter et al., 1990). Should this strategy fail, then participants may proceed to a guessing strategy, which may still be in part related to their cognitive ability (Vodegel Matzen et al., 1994). Thus, it appears that the responding process of Raven's matrices corresponds to the description of the processes that are best modeled with Nested Logit Models (Suh & Bolt, 2010): Examinees are expected to find the correct answer through the identification and application of the rule of the matrix (Level 1). Should the examinee not be able to properly identify or apply the rule, then and only then would the distractors be considered as potential responses (Level 2).

Therefore, we hypothesized that the NLM would better fit the responding process than the NRM, and would thus outperform it here. More specifically, a NLM with a 3PL level 1 model – in other words, a 3PNL (for 3-Parameter Nested Logistic) model – was hypothesized to fit the nominal data the best, for the same reasons (previously explained) that we hypothesized the 3PL model to best fit the dichotomized data.

Because binary and polytomous models do not fit the same variables, they cannot be compared with typical IRT model comparison tools – such as Likelihood Ratio Tests. However, the capability of polytomous IRT to recover supplementary information from the distractors may help result in slightly different, and potentially more reliable estimates of θ, especially for test takers of low ability (Bock, 1972). For this reason, we hypothesized that the best fitting polytomous model (hypothetically the 3PNL model) would outperform the best fitting dichotomous model (hypothetically the 3PL model) in terms of test reliability.

2. Method

2.1. Participants

The sample was composed of a total of 499 undergraduate students of a French business school. All participants – 214 males and 285 females, aged between 19 and 24 ($M = 20.7$, $SD = 0.93$) – were French speakers. They responded the SPM-LS individually and on a voluntary basis. The SPM-LS and basic demographic questions were all presented on computer. The participants received no compensation for participation, and had received no previous introduction to or training in intelligence research or psychometrics.

2.2. Instrument

The SPM-LS is composed of the last series of the Standard Progressive Matrices (Raven, 1941). This last series is the series of the test has the expected highest overall difficulty of the SPM, and it is composed of 12 items of increasing difficulty. Each item is an incomplete 3 × 3 matrix of non-verbal stimuli, which are related by logical rules. The last of the 9 stimuli is left blank, and the examinee is to identify the missing stimulus among 8 possible answers – 1 correct and 7 distractors. To do so, the participant has to be able to identify the logical rule used in the incomplete matrix, and to apply that rule to identify the missing stimulus. The participants had no time limit to answer the 12 items, and were encouraged to respond every item, even when they were unsure of their response.

2.3. Data analysis

2.3.1. CTT analyses

We noted earlier that the other short versions of Raven's matrices (Arthur & Day, 1994; Bors & Stokes, 1998) have relied on traditional CTT methods. Thus, for the sake of comparability between the SPM-LS and the short APM versions, similar CTT-based methods were used on the SPM-LS.

First, to explore the dimensionality of the SPM-LS, based on the tetrachoric correlations between the 12 items, we conducted an Exploratory Factor Analysis (EFA) with parallel analysis (Hayton, Allen, & Scarpello, 2004; Horn, 1965), as implemented in the R package 'psych' (Revelle, 2017).

The SPM-LS being theoretically unidimensional, we also used Confirmatory Factor Analysis (CFA) to examine the fit of a unidimensional model. The CFA was performed with the R package 'lavaan' (Rosseel, 2012), with Weighted Least Squares Means and Variance adjusted (WLSMV) estimation. As typically recommended (Hu & Bentler, 1999; Yu, 2002), we used the Comparative Fit Index (CFI) and Tucker-Lewis Index (TLI) with cut-offs of 0.95, the Standardized Root Mean Square Residual (SRMR) with a cut-off of 0.08, the Root Mean Square Error of Approximation (RMSEA) with a cut-off of 0.06, and the Weighted Root Mean Square Residual (WRMR), with a cut-off of 1.0.

For reliability, we computed Cronbach's α, as well as the now recommended unidimensionality index McDonald's ω_h (Revelle & Zinbarg, 2009; Zinbarg, Yovel, Revelle, & McDonald, 2006). McDonald's ω_h was based on the CFA model and computed with the package 'semTools' (Contributors, 2016).

2.3.2. Binary IRT models

1-4PL models were fitted using the R package 'mirt' (Chalmers, 2012). The models were compared using various methods, including the corrected Akaike Information Criterion (AICc) – a lower statistic representing a better fit – Likelihood Ratio Tests, and M_2–based indicators of Goodness-of-Fit (Maydeu-Olivares, 2013). These indicators included the Comparative Fit Index (CFI) and Tucker-Lewis Index (TLI) with cut-offs of 0.95, the Standardized Root Mean Square Residual (SRMR) with a cut-off of 0.08, and the Root Mean Square Error of Approximation (RMSEA) with a cut-off of 0.06.

In addition, we computed the estimates of empirical reliability from the IRT test information functions (Raju, Price, Oshima, & Nering, 2007). Empirical reliability corresponds to the expected reliability in the sample distribution of the θ scores, as estimated through the IRT models. We also computed marginal reliability estimates, which are typically used in conjunction with empirical reliability estimates (e.g., Myszkowski & Storme, 2017), and correspond to the expected reliability in an (assumed) normal prior density of θ.

2.3.3. Polytomous IRT models

Polytomous models of the correct responses and the distractor responses have the disadvantage of being considerably more parametrized than binary models – for the reason that they model here 8 probabilities per item, while binary IRT models only model two. Therefore, although a unique convergent solution was still found for the nominal models, M_2 –based indicators of Goodness-of-Fit failed to be computed because of the lack of degrees of freedom. However, the nominal models could still be compared with one another using the corrected Akaike Information Criterion (AICc), and, for the nested models, with Likelihood Ratio Tests (LRT). Additionally, we compared the 3PNL models with their binary counterparts (3PL). Because they do not fit the same manifest variables, their fit could not be compared, but

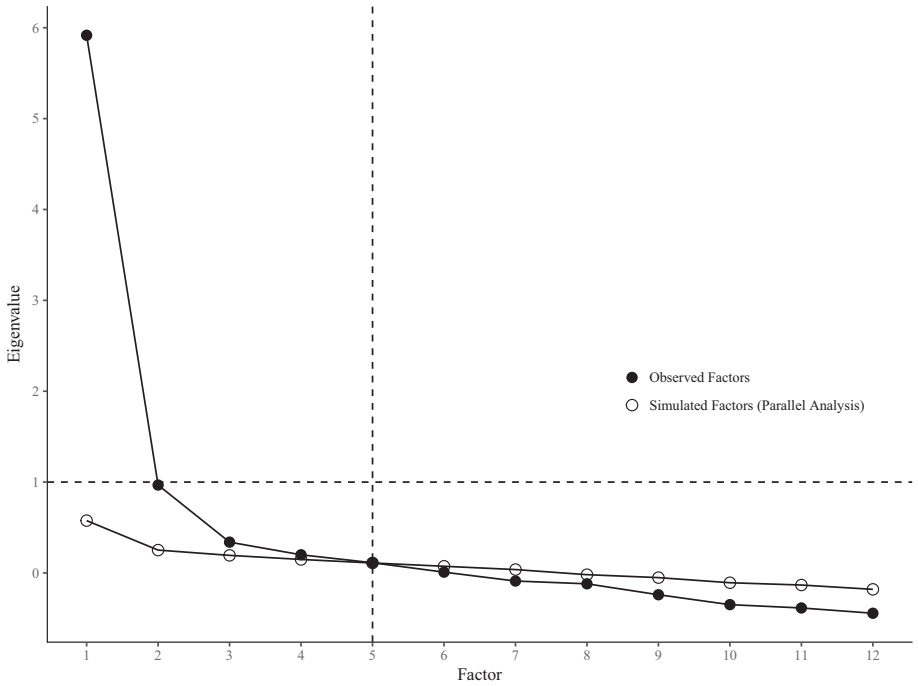

Fig. 1. Scree plot of the exploratory factor analysis (with parallel analysis).

their estimates of empirical and marginal reliability can be (Bock, 1972). To infer on reliability gains, in the absence of a prescribed test, we computed basic bootstrapped confidence intervals – based on case resampling – and bootstrapped Wald's z tests of the reliability gains. Finally, to compare the scoring strategies resulting from the different models, we inspected scatterplots of the relations between sum scores, CFA factor scores, and the factor scores of all IRT models using JASP (JASP Team, 2018).

3. Results

3.1. CTT analyses

The EFA produced mixed results in terms of dimensionality: The Parallel Analysis indicated to retain 5 factors, while only the first factor had an eigenvalue above 1. Further, as can be seen on the scree plot presented in Fig. 1, a drop in eigenvalues was observed between the first factor (with an eigenvalue of 5.92) and the other factors (eigenvalues of 0.97 and below). Finally, the loadings on factors 2 to 5 did not appear interpretable. The unidimensional model tested through CFA indicated a borderline acceptable structural validity (CFI = 0.945, TLI = 0.933, RMSEA = 0.079, SRMR = 0.108, WRMR = 1.50). Moreover, the SPM-LS had satisfactory reliability estimates, with a Cronbach's α of 0.92, and a McDonald's unidimensionality estimate of ω_h of 0.86. Overall, although not meeting every expectation, these indices compare advantageously to the indices that have been reported for the short APM (Arthur & Day, 1994; Bors & Stokes, 1998). Further, these results lead us to proceed to IRT analyzes assuming the structure of the SPM-LS to be essentially unidimensional.

3.2. Binary IRT analyses

The absolute fit indices for the IRT binary models, as well as their estimates of empirical and marginal reliability, are reported in Table 1.

As hypothesized, the 3PL and 4PL had an satisfactory fit to the data. Further, the 3PL model had the best fit, in both the goodness-of-fit indices and the Likelihood Ratio Tests – the 3PL model presenting a significantly better fit that the 1PL (χ^2 = 182.653, df = 23, p < .001) and the 2PL (χ^2 = 69.979, df = 12, p < .001) models. The Likelihood Ratio Test suggested that the 4PL fit only marginally better than the 3PL (χ^2 = 20.909, df = 12, p = .052). A lower AIC for the 3PL led us to further use the 3PL as the best fitting model, but it should be noted that overall, the differences in goodness of fit between the 3PL and 4PL were very minimal.

As hypothesized, the SPM-LS with showed satisfactory empirical reliability (empirical $r_{xx',3PL}$ = 0.843, Bootstrapped 95% CI [0.828, 0.861]). The marginal reliability estimate for the 3PL model indicates that observations from an assumed normal prior distribution produce overall reliable estimates (marginal $r_{xx',3PL}$ = 0.829, Bootstrapped 95% CI [0.812, 0.841]).

3.3. Polytomous IRT analyses

The tested polytomous models being too heavily parametrized, their M_2 goodness of fit indices could not be computed. They could however be compared with each other through Likelihood Ratio Tests. As hypothesized, the 3PNL model appeared to fit the best the nominal data, with a significantly better fit than the Nominal Response Model (χ^2 = 142.333, df = 12, p < .001) and the 2PNL model (χ^2 = 65.295, df = 12, p < .001). The 4PNL model did not fit the data significantly better than the 3PNL model (χ^2 = 19.410, df = 12, p = .08). The comparison of AICc for the different models also supported this conclusion. The Item Characteristic Curves of the 3PNL model are presented on Fig. 2.

The 3PL and the 3PNL models being the best fitting models for respectively the dichotomous and the nominal responses, we compared their reliability estimates to compare the two approaches. As hypothesized, the information from distractors recovered by the 3PNL model allows to obtain more reliable θ estimates in the low ability domain, with an empirical reliability of 0.864 (Bootstrapped 95% CI [0.851, 0.889]), and a marginal reliability of 0.853 (Bootstrapped 95% CI [0.843, 0.873]). A significant gain in reliability – both empirical ($\Delta r_{xx'}$ = 0.029, Bootstrapped 95% CI [0.019, 0.036], z = 6.06, p < .001) and marginal ($\Delta r_{xx'}$ = 0.024, Bootstrapped 95% CI [0.020, 0.041], z = 5.63, p < .001) – was achieved through the use of the 3PNL model. The plot presented in Fig. 3 presents the reliability

Table 1
Dichotomous IRT fit indices and empirical reliability of the SPM-LS.

Model	χ^2	df	CFI	TLI	SRMR	RMSEA	AICc	Empirical reliability	Marginal reliability
1PL (Rasch)	346.061	65	0.899	0.898	0.106	0.093	5723.0	0.797	0.626
2PL	226.291	54	0.938	0.925	0.065	0.080	5634.1	0.805	0.792
3PL	114.266	42	0.974	0.959	0.056	0.059	5591.3	0.843	0.829
4PL	100.085	30	0.975	0.945	0.051	0.068	5599.1	0.832	0.779

Note. CFI, Comparative Fit Index; TLI, Tucker-Lewis Index; SRMR, Standardized Root Mean Square Residual; RMSEA, Root Mean Square Error of Approximation; AICc, Akaike Information Criterion (corrected).

differences of the two approaches as a function of θ. It shows that the gain in reliability is essentially found in low to average ability – in reference to our sample.

Related to this point, we can observe on Fig. 4, which presents the correlations between the various scoring strategies, that, in spite of nearly perfect correlations between the scores of the different scoring methods, the Nested Logit Models provide different estimations that the other methods in the low ability levels.

4. Discussion

First investigated with CTT-based methods for the sake of comparability, the reliability indices and fit indices of a unidimensional CFA compared advantageously to the indices that have been previously reported for the short APM (Arthur & Day, 1994; Bors & Stokes, 1998). While we argue that these investigations are not appropriate models for the responses at the SPM-LS, they still suggest that the SPM-LS presents better psychometrical qualities in than the short APMs, regardless of the psychometrical framework used.

Investigated through IRT models, the SPM-LS presented adequate structural validity, and a strong reliability – outperforming reliability and structural validity estimates previously reported for short forms of the Advanced Progressive Matrices (Arthur et al., 1999; Bors & Stokes, 1998).

IRT modeling advances offer interesting possibilities in the investigation and scoring of ability measures. An important advance for the modeling of responses to multiple-choice items – which the SPM and APM are composed of – is the possibility to recover information from distractor responses through the modeling of the full polytomous responses (Bock, 1972), rather than the dichotomized pass-fail responses. The literature highlights two modeling approaches corresponding to different problem solving processes: Nominal Response Models (Bock, 1972) – in which distractors responses and the correct response are modeled on the same level, and thus considered as all competing in the solving process – and Nested Logit Models (Suh & Bolt, 2010) – in which responses of distractors are modeled separately from the correct response, representing situations where the distractors are considered only when the initial problem-solving strategy failed.

In the case of Raven's matrices, both are certainly conceivable: A respondent could either compare all potential responses directly, based on their "fit" with the matrix – a process for which the NRM would be appropriate – or start by trying to extract a rule in the matrix and apply it to find the correct response, and, should this fail, proceed to a comparison strategy – a process for which the Nested Logit Models

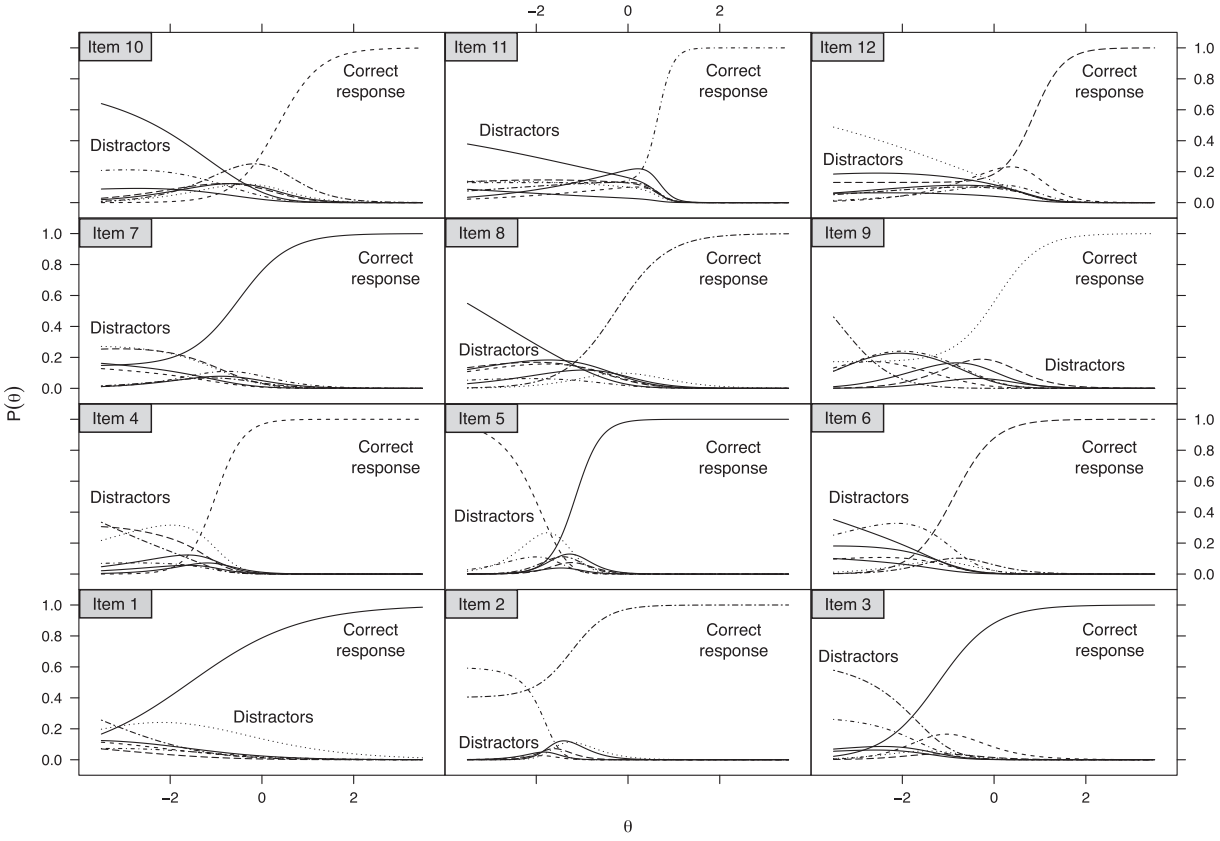

Fig. 2. Item characteristic curves of the best fitting (3PNL) model.

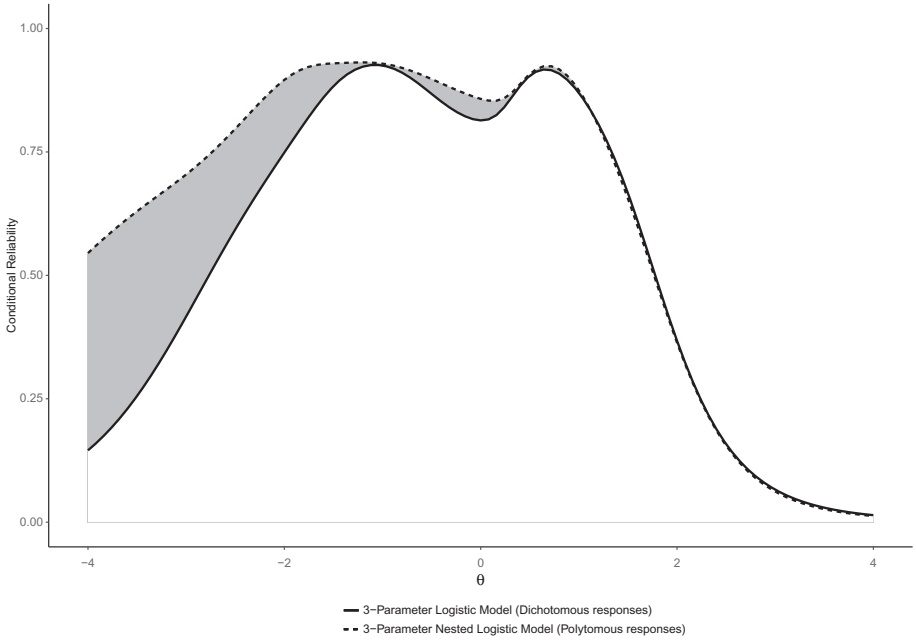

Fig. 3. Gain in reliability from recovering distractor response information.

would be appropriate. In any case, the results of the present study notably suggest that, in multiple choice tests like Raven's matrices, pseudo-guessing is a phenomenon that should be accounted for. Further, the incremental reliability resulting from recovering distractor information through the Nested Logit Models indicates that distractor responses bring information about the latent ability that can be useful to at least attempt recovering. It further indicates that one may consider some distractor responses as "partially correct", in the sense that some incorrect responses are more indicative of higher ability than others.

In the case of the SPM-LS, the recovery of information resulting from the use of Nested Logit Models, albeit significant, may appear minor, but a closer look at the reliability functions presented in Fig. 3 actually indicates important gains in the low ability domains. As an important note, we mean "low ability" as compared with the average level of our sample of post-secondary students, thus "low abilities" here would likely be closer to average abilities in the general population. We thus strongly recommend, sample size permitting, the use of these modeling (and scoring) approaches, especially for populations that are not expected to have overall high performances at this test. More generally, we believe that future investigations of similar tests (including the SPM, the APM and their short versions) should consider such methods, which account for important features of the psychological testing situation and result in reliability gains.

4.1. Limitations

This study certainly has limitations. First and foremost, the convenience sample used is not an accurate representation of the general population. In particular, although we advanced that the SPM still captures variability in intelligence among post-secondary students, our reliability analysis shows that the test is probably too easy for the population tested – above 2 standard deviations above the mean, the reliability drops. Further, 52 out of 499 (approximately 10.4%) of the sample had a perfect score of 12, suggesting a ceiling effect. This result can be contrasted with research indicating that the full SPM discriminates up to the 99% IQ percentile (Jensen et al., 1988). Overall, these results suggest that the SPM-LS is probably more appropriate for populations of lower expected intelligence levels than our sample. Furthermore, the estimates presented in this study are dependent upon the sample, and the psychometric qualities observed in this sample need further replication. For these reasons, we recommend that this study be replicated on a variety of samples, using other sampling strategies – especially random sampling from the general population.

In addition, this research suggests that the dimensionality of the SPM-LS should be further investigated (or strengthened). Indeed, although the results of the EFA and CFA indicated mixed results, and suggest that the SPM-LS may not be a purely unidimensional measure. We recommend that the SPM-LS be investigated in larger samples to further the investigation of its dimensionality.

Also, it is important to point out that only internal reliability and factor structure were investigated in this study. We thus suggest that further studies investigate other important psychometrical qualities such as test-retest reliability and concurrent validity – which was done for the short APMs. So far, the concurrent validity and test-retest reliability of the SPM-LS can only be inferred from those of the original SPM, which is obviously insufficient. Related to this point, a limitation of this study is the lack of a comparison between the full length SPM and the SPM-LS. The SPM-LS was here investigated as a standalone test, which prevents the SPM-LS responses analyzed from being potentially contaminated by responding the first 4 series of the SPM, but also prevents an actual comparison between the two tests. Thus, it is impossible from this study to discuss the loss in test information that results from shortening the SPM.

On a more technical note, the calibration of the Nested Logit Models was substantially longer and required a much larger number of iterations to converge than the binary models and the Nominal Response Model. Overcoming these challenges is of course dependent upon various factors (sample size, the number of distractors, the use of priors, the optimizer used, etc.), but researchers interested in applying Nested Logit Models should be aware that such models may not be computationally parsimonious. Related to this, the calibrated Nested Logit Models being heavily parametrized, their parameter estimates may be very sensitive to the sample and thus less stable than the estimates of less complex models (for example binary models). Future studies may examine the reproducibility of the parameters of Nested Logit Models through cross-validation methods.

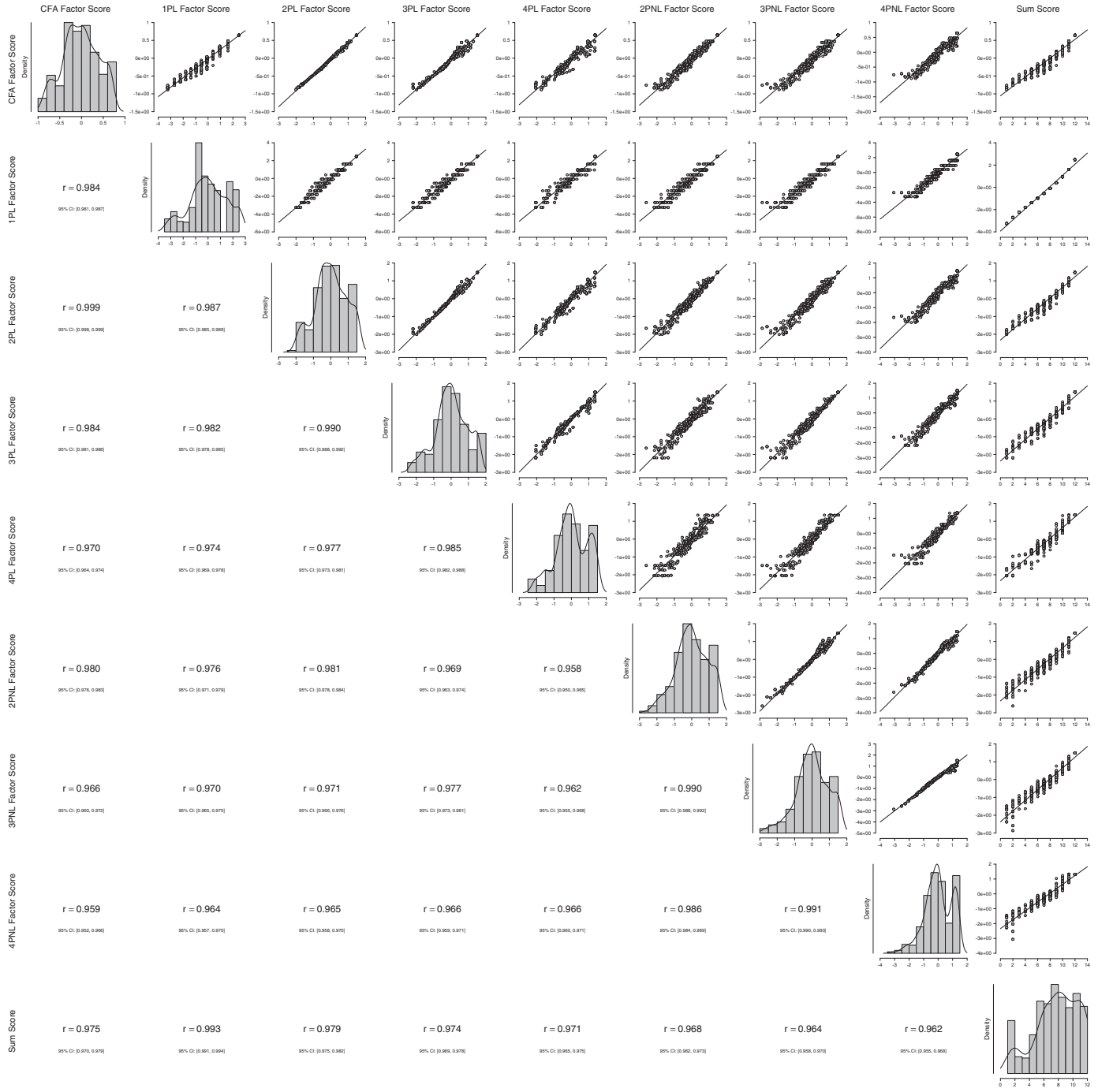

Fig. 4. Scatterplot matrix with density plots.

5. Conclusion

To conclude, the SPM-LS appeared in this study as a shorter alternative to the SPM and as a more robust alternative to the short APMs. While this study certainly calls for further investigations focused on other qualities and other populations, we would suggest using the SPM-LS in populations that share similarities with our sample, notably convenience samples of post-secondary students (a strategy commonly used in psychological research) similar to the present one, as well as samples expected to have lower average *g* levels. Even though the various scoring strategies used appeared to strongly correlate, we recommend scoring the SPM-LS through 3-4PL Item Response Theory models. Further, although binary models already present a good fit to the data collected with the test, we recommend – provided sufficient sample size – the use of Nested Logit Models to recover information from the distractor responses and increase scoring reliability for low abilities.

References

Ablard, K. E., & Mills, C. J. (1996). Evaluating abridged versions of the Raven's Advanced Progressive Matrices for identifying students with academic talent. *Journal of Psychoeducational Assessment, 14*(1), 54–64. http://dx.doi.org/10.1177/073428299601400105.

Arthur, W., & Day, D. V. (1994). Development of a short form for the Raven Advanced Progressive Matrices test. *Educational and Psychological Measurement, 54*(2), 394–403. http://dx.doi.org/10.1177/0013164494054002013.

Arthur, W., Tubre, T. C., Paul, D. S., & Sanchez-Ku, M. L. (1999). College-sample

psychometric and normative data on a short form of the Raven Advanced Progressive Matrices test. *Journal of Psychoeducational Assessment, 17*(4), 354–361. http://dx.doi.org/10.1177/073428299901700405.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*(1), 29–51. http://dx.doi.org/10.1007/BF02291411.

Bors, D. A., & Stokes, T. L. (1998). Raven's Advanced Progressive Matrices: Norms for first-year university students and the development of a short form. *Educational and Psychological Measurement, 58*(3), 382–398. http://dx.doi.org/10.1177/0013164498058003002.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika, 71*(3), 425–440. http://dx.doi.org/10.1007/s11336-006-1447-6.

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices test. *Psychological Review, 97*(3), 404–431. http://dx.doi.org/10.1037/0033-295X.97.3.404.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1–29.

Contributors (2016). semTools: Useful tools for structural equation modeling. Retrieved from https://CRAN.R-project.org/package=semTools.

Dillon, R. F., Pohlmann, J. T., & Lohman, D. F. (1981). A factor analysis of Raven's Advanced Progressive Matrices freed of difficulty factors. *Educational and Psychological Measurement, 41*(4), 1295–1302. http://dx.doi.org/10.1177/001316448104100438.

Gignac, G. E. (2015). Raven's is not a pure measure of general intelligence: Implications for g factor theory and the brief measurement of g. *Intelligence, 52*, 71–79. http://dx.doi.org/10.1016/j.intell.2015.07.006.

Hayton, J. C., Allen, D. G., & Scarpello, V. (2004). Factor retention decisions in Exploratory Factor Analysis: A tutorial on Parallel Analysis. *Organizational Research Methods, 7*(2), 191–205. http://dx.doi.org/10.1177/1094428104263675.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*(2), 179–185. http://dx.doi.org/10.1007/BF02289447.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1–55. http://dx.doi.org/10.1080/10705519909540118.

JASP Team (2018). JASP (Version 0.8.5) [Computer software]. Retrieved from https://jasp-stats.org/.

Jensen, A. R., Saccuzzo, D. P., & Larson, G. E. (1988). Equating the standard and advanced forms of the Raven Progressive Matrices. *Educational and Psychological Measurement, 48*(4), 1091–1095. http://dx.doi.org/10.1177/0013164488484026.

Lynn, R., Allik, J., & Irwing, P. (2004). Sex differences on three factors identified in Raven's standard progressive matrices. *Intelligence, 32*(4), 411–424. http://dx.doi.org/10.1016/j.intell.2004.06.007.

Macdonald, P., & Paunonen, S. V. (2002). A Monte Carlo comparison of item and person statistics based on Item Response Theory versus Classical Test Theory. *Educational and Psychological Measurement, 62*(6), 921–943. http://dx.doi.org/10.1177/0013164402238082.

Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of Item Response Theory models. *Measurement: Interdisciplinary Research and Perspectives, 11*(3), 71–101. http://dx.doi.org/10.1080/15366367.2013.831680.

Myszkowski, N., & Storme, M. (2017). Measuring "good taste" with the Visual Aesthetic Sensitivity Test-Revised (VAST-R). *Personality and Individual Differences, 117*, 91–100. http://dx.doi.org/10.1016/j.paid.2017.05.041.

Myszkowski, N., Storme, M., Zenasni, F., & Lubart, T. (2014). Is visual aesthetic sensitivity independent from intelligence, personality and creativity? *Personality and Individual Differences, 59*, 16–20. http://dx.doi.org/10.1016/j.paid.2013.10.021.

Pind, J., Gunnarsdóttir, E. K., & Jóhannesson, H. S. (2003). Raven's standard progressive matrices: New school age norms and a study of the test's validity. *Personality and Individual Differences, 34*(3), 375–386. http://dx.doi.org/10.1016/S0191-8869(02)00058-2.

Raju, N. S., Price, L. R., Oshima, T. C., & Nering, M. L. (2007). Standardized conditional SEM: A case for conditional reliability. *Applied Psychological Measurement, 31*(3), 169–180. http://dx.doi.org/10.1177/0146621606291569.

Raven, J. C. (1941). Standardization of Progressive Matrices, 1938. *British Journal of Medical Psychology, 19*(1), 137–150. http://dx.doi.org/10.1111/j.2044-8341.1941.tb00316.x.

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

Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika, 74*(1), 145–154. http://dx.doi.org/10.1007/s11336-008-9102-z.

Rosseel, Y. (2012). lavaan: An R package for Structural Equation Modeling. *Journal of Statistical Software, 48*(2), 1–36.

Suh, Y., & Bolt, D. M. (2010). Nested logit models for multiple-choice item response data. *Psychometrika, 75*(3), 454–473. http://dx.doi.org/10.1007/s11336-010-9163-7.

Vodegel Matzen, L. B. L., van der Molen, M. W., & Dudink, A. C. M. (1994). Error analysis of Raven test performance. *Personality and Individual Differences, 16*(3), 433–445. http://dx.doi.org/10.1016/0191-8869(94)90070-1.

Yu, C.-Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes.* Los Angeles: University of California.

Zinbarg, R. E., Yovel, I., Revelle, W., & McDonald, R. P. (2006). Estimating generalizability to a latent variable common to all of a Scale's indicators: A comparison of estimators for ω_h. *Applied Psychological Measurement, 30*(2), 121–144. http://dx.doi.org/10.1177/0146621605278814.