Journal of
Intelligence

Analysis of an Intelligence Dataset

Edited by

Nils Myszkowski

Printed Edition of the Special Issue Published in *Journal of Intelligence*

www.mdpi.com/journal/jintelligence

MDPI

Analysis of an Intelligence Dataset

Analysis of an Intelligence Dataset

Editor

Nils Myszkowski

MDPI • Basel • Beijing • Wuhan • Barcelona • Belgrade • Manchester • Tokyo • Cluj • Tianjin

Editor
Nils Myszkowski
Department of Psychology,
Pace University
USA

Editorial Office
MDPI
St. Alban-Anlage 66
4052 Basel, Switzerland

This is a reprint of articles from the Special Issue published online in the open access journal *Journal of Intelligence* (ISSN 2079-3200) (available at: https://www.mdpi.com/journal/jintelligence/special_issues/intelligence_dataset).

For citation purposes, cite each article independently as indicated on the article page online and as indicated below:

LastName, A.A.; LastName, B.B.; LastName, C.C. Article Title. *Journal Name* **Year**, *Volume Number*, Page Range.

ISBN 978-3-0365-0040-9 (Hbk)
ISBN 978-3-0365-0041-6 (PDF)

Contents

About the Editor

Nils Myszkowski is an Assistant Professor of Psychology at Pace University (NYC). A graduate of Paris Descartes University (Paris, France), his central research interest is the application and improvement of psychometric methods to measure and understand intellectual/emotional/creative/aesthetic abilities, especially applied to occupational contexts. He has authored or coauthored over 30 peer-reviewed journal articles, developed 3 packages for R, and received in 2020 the Daniel E. Berlyne Award in recognition of outstanding contributions by an early-career scholar from the American Psychological Association (Division 10: Society for the Psychology of Aesthetics, Creativity and the Arts).

Journal of
Intelligence

MDPI

Editorial

Analysis of an Intelligence Dataset

Nils Myszkowski

Department of Psychology, Pace University, New York, NY 10038, USA; nmyszkowski@pace.edu

Received: 28 October 2020; Accepted: 4 November 2020; Published: 19 November 2020

It is perhaps popular belief—at least among non-psychometricians—that there is a unique or standard way to investigate the psychometric qualities of tests. If anything, the present Special Issue demonstrates that it is not the case. On the contrary, this Special Issue on the "analysis of an intelligence dataset" is, in my opinion, a window to the present vividness of the field of psychometrics.

Much like an invitation to revisit a story with various styles or with various points of view, this Special Issue was opened to contributions that offered extensions or reanalyses of a single—and somewhat simple—dataset, which had been recently published. The dataset was from a recent paper (Myszkowski and Storme 2018), and contained responses from 499 adults to a non-verbal logical reasoning multiple-choice test, the SPM–LS, which consists of the Last Series of Raven's Standard Progressive Matrices (Raven 1941). The SPM–LS is further discussed in the original paper (as well as through the investigations presented in this Special Issue), and most researchers in the field are likely familiar with the Standard Progressive Matrices. The SPM–LS is simply a proposition to use the last series of the test as a standalone test. A minimal description of the SPM–LS would probably characterize it as a theoretically unidimensional measure—in the sense that one ability is tentatively measured—comprised of 12 pass-fail non-verbal items of (tentatively) increasing difficulty. Here, I refer to the pass-fail responses as the binary responses, and the full responses (including which distractor was selected) as the polytomous responses. In the original paper, a number of analyses had been used, including exploratory factor analysis with parallel analysis, confirmatory factor analyses using a structural equation modeling framework, binary logistic item response theory models (1-, 2-, 3- and 4- parameter models), and polytomous (unordered) item response theory models, including the nominal response model (Bock 1972) and nested logit models (Suh and Bolt 2010). In spite of how extensive the original analysis may have seemed, the contributions of this Special Issue present several extensions to our analyses.

I will now briefly introduce the different contributions of the Special Issue, in chronoligical order of publication. In their paper, Garcia-Garzon et al. (2019) propose an extensive reanalysis of the dimensionality of the SPM–LS, using a large variety of techniques, including bifactor models and exploratory graph analysis. Storme et al. (2019) later find that the reliability boosting strategy proposed in the original paper—which consisted of using nested logit models (Suh and Bolt 2010) to recover information from distractor information—is useful in other contexts, by using the example on a logical reasoning test applied in a personnel selection context. Moreover, Bürkner (2020) later presents how to use his R Bayesian multilevel modeling package `brms` (Bürkner 2017) in order to estimate various binary item response theory models, and compares the results with the frequentist approach used in the original paper with the item response theory package `mirt` (Chalmers 2012). Furthermore, Forthmann et al. (2020) later proposed a new procedure that can be used to detect (or select) items that could present discriminating distractors (i.e., items for which distractor responses could be used to extract additional information). In addition, Partchev (2020) then discusses issues that relate to the use of distractor information to extract information on ability in multiple choice tests, in particular in the context of cognitive assessment, and presents how to use the R package `dexter` (Maris et al. 2020) to study the binary responses and distractors of the SPM–LS.

I then present an analysis of the SPM–LS (especially of its monotonicity) using (mostly) the framework of Mokken scale analysis (Mokken 1971). Finally, Robitzsch (2020) proposes new procedures for latent class analysis applied on the polytomous responses, combined with regularization to obtain models of parsimonious complexity.

It is interesting to note that, in spite of the relative straightforwardness of the task and the relative simplicity of the dataset—which in the end, contains answers to a few pass-fail items in a (theoretically) unidimensional instrument—the contributions of this Special Issue offer a lot of original and new perspectives on analyzing intelligence test data. Admittedly, much like the story retold 99 times in Queneau's *Exercices de style*, the dataset reanalysed in this Special Issue is, in of itself, of moderate interest. Nevertheless, the variety, breadth and complementarity of the procedures used, proposed and described here clearly demonstrate the creative nature of the field, giving an echo to the proposition by Thissen (2001) to see artistic value in psychometric engineering. I would like to thank Paul De Boeck for proposing the topic of this Special Issue and inviting me to act as guest editor, as well as the authors and reviewers of the articles published in this issue for their excellent contributions. I hope that the readers of *Journal of Intelligence* will find as much interest in them as I do.

Funding: This research received no external funding.

Conflicts of Interest: The author declares no conflict of interest.

References

Bock, R. Darrell. 1972. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika* 37: 29–51. [CrossRef]

Bürkner, Paul-Christian. 2017. Brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software* 80: 1–28. [CrossRef]

Bürkner, Paul-Christian. 2020. Analysing Standard Progressive Matrices (SPM-LS) with Bayesian Item Response Models. *Journal of Intelligence* 8: 5. [CrossRef]

Chalmers, R. Philip. 2012. Mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software* 48: 1–29. [CrossRef]

Forthmann, Boris, Birgit Schütze Natalie Förster, Karin Hebbecker, Janis Flessner, Martin T. Peters, and Elmar Souvignier. 2020. How Much g Is in the Distractor? Re-Thinking Item-Analysis of Multiple-Choice Items. *Journal of Intelligence* 8: 11. [CrossRef] [PubMed]

Garcia-Garzon, Eduardo, Francisco J. Abad, and Luis E. Garrido. 2019. Searching for G: A New Evaluation of SPM-LS Dimensionality. *Journal of Intelligence* 7: 14. [CrossRef] [PubMed]

Maris, Gunter, Timo Bechger, Jesse Koops, and Ivailo Partchev. 2020. dexter: Data Management and Analysis of Tests. Available online: https://rdrr.io/cran/dexter/ (accessed on 6 November 2020).

Mokken, Robert J. 1971. *A Theory and Procedure of Scale Analysis.* The Hague and Berlin: Mouton/De Gruyter.

Myszkowski, Nils, and Martin Storme. 2018. A snapshot of g? Binary and polytomous item-response theory investigations of the last series of the Standard Progressive Matrices (SPM-LS). *Intelligence* 68: 109–16. [CrossRef]

Partchev, Ivailo. 2020. Diagnosing a 12-Item Dataset of Raven Matrices: With Dexter. *Journal of Intelligence* 8: 21. [CrossRef] [PubMed]

Raven, John C. 1941. Standardization of Progressive Matrices, 1938. *British Journal of Medical Psychology* 19: 137–50. [CrossRef]

Robitzsch, Alexander. 2020. Regularized Latent Class Analysis for Polytomous Item Responses: An Application to SPM-LS Data. *Journal of Intelligence* 8: 30. [CrossRef] [PubMed]

Storme, Martin, Nils Myszkowski, Simon Baron, and David Bernard. 2019. Same Test, Better Scores: Boosting the Reliability of Short Online Intelligence Recruitment Tests with Nested Logit Item Response Theory Models. *Journal of Intelligence* 7: 17. [CrossRef] [PubMed]

Suh, Youngsuk, and Daniel M. Bolt. 2010. Nested Logit Models for Multiple-Choice Item Response Data. *Psychometrika* 75: 454–73. [CrossRef]

Thissen, David. 2001. Psychometric engineering as art. *Psychometrika* 66: 473–85. [CrossRef]

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Journal of *Intelligence*

Article

Searching for G: A New Evaluation of SPM-LS Dimensionality

Eduardo Garcia-Garzon [1,*], Francisco J. Abad [1] and Luis E. Garrido [2]

[1] Facultad de Psicología, Universidad Autónoma de Madrid, 28049 Madrid, Spain
[2] Facultad de Psicología, Pontificia Universidad Católica Madre y Maestra,
Santo Domingo 10109, Dominican Republic
[*] Correspondence: eduardo.garciag@uam.es; Tel.: +34-91-497-8750

Received: 26 April 2019; Accepted: 25 June 2019; Published: 28 June 2019

Abstract: There has been increased interest in assessing the quality and usefulness of short versions of the Raven's Progressive Matrices. A recent proposal, composed of the last twelve matrices of the Standard Progressive Matrices (SPM-LS), has been depicted as a valid measure of *g*. Nonetheless, the results provided in the initial validation questioned the assumption of essential unidimensionality for SPM-LS scores. We tested this hypothesis through two different statistical techniques. Firstly, we applied exploratory graph analysis to assess SPM-LS dimensionality. Secondly, exploratory bi-factor modelling was employed to understand the extent that potential specific factors represent significant sources of variance after a general factor has been considered. Results evidenced that if modelled appropriately, SPM-LS scores are essentially unidimensional, and that constitute a reliable measure of *g*. However, an additional specific factor was systematically identified for the last six items of the test. The implications of such findings for future work on the SPM-LS are discussed.

Keywords: Raven matrices; Standard Progressive Matrices test; dimensionality; bi-factor; parallel analysis; target rotation; exploratory graph analysis

1. Introduction

The Standard Progressive Matrices (i.e., SPM [1]), in any of its forms, constitutes one of the most applied tests for measuring general intelligence (*g*). Due to its considerable length (60 items), there has been a growing interest in developing short versions of this test. Unfortunately, the available short versions—such as the Advanced Progressive Matrices tests (i.e., APM)—present substantial shortcomings [2]. Consequently, [2] proposed the SPM-LS, a new short version of the SPM test based on its last, most-difficult 12 matrices of this test. These items consist of non-verbal stimuli where each item presents a single correct answer and seven distractors. In its recent validation, the SPM-LS scores were analysed using exploratory and confirmatory factor analyses as well as item response theory models as follows: After concluding that the SPM-LS scores were sufficiently unidimensional, individual responses were modelled with the 1 to 4 parameter logistic models. Additionally, a three-parameter nested logistic model was applied to recover relevant information from responses to the different distractors. Remarkably, the original authors concluded that the SPM-LS was a superior alternative to the APM test ([2]; p.113), and encouraged other researchers to re-analyse this dataset by making it publicly available and by opening a call for papers on the matter in the Journal of Intelligence.

As part of this call, this investigation will re-evaluate [2] claim of SPM-LS being essentially unidimensional. This claim is vital to understand if SPM-LS represents a valid measure of *g* and represent a necessary assumption for many of the following analysis presented by the original authors. As [2] acknowledged that "SPM-LS may not be a purely unidimensional measure" (p.114), we decided to analyse SPM-LS dimensionality by expanding the original approaches with the application of network-based exploratory analysis and bi-factor modelling.

1.1. On the Progressive Matrices Dimensionality

Few consensuses are more extended in the intelligence literature than the belief that the SPM test [1] represents a consistent measure of general intelligence (*g*; Panel A, Figure 1). Even though this claim has received overwhelming support in the literature [3–5], other authors have considered general intelligence to be a broader construct to be measured with different tasks and item formats [6]. Be that as it may, support for strict unidimensionality has historically been equivocal for short SMP versions such as the APM test. As early as 1981, some authors found evidence of an orthogonal two-factor model [7,8] were among the first authors to suggest that a nuisance factor, corresponding to a "speed factor", could be found for APM scores (Panel C, Figure 1). [3] found that the two-factor proposed in [2] fitted the data better than the single factor model if the inter-factor correlation was estimated. Nevertheless, the high magnitude of this correlation (i.e., 0.89; Panel B, Figure 1; [3]), in conjunction with the inspection of fit statistics, was taken as evidence in favour of a unidimensional model. Since then, other authors on the field have supported [3] conclusions [4,5].

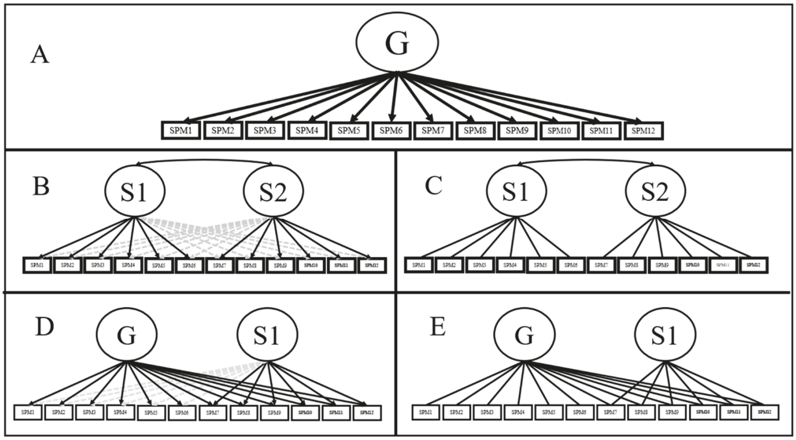


Figure 1. Schematic representation of theoretical SPM-LS models: (**A**): Unidimensional model; (**B**) Exploratory bi-dimensional model; (**C**): Confirmatory bi-dimensional model; (**D**): Exploratory bi-factor model; (**E**): Confirmatory bi-factor model. Arrows in black represent estimated paths for CFA models, and untargeted loadings in EFA models. Grey arrows represent targeted (minimised) loadings during EFA target rotation.

Recent applications of bi-factor modelling offered new insights regarding the dimensionality of the APM, as well as the role of potential secondary factors (Panel E, Figure 1). As the bi-factor model simultaneously estimates a general plus several orthogonal specific factors [9], it provides a clear separation of such different sources of variation. Noteworthy, as specific factors only account for a variance that is residual to the general factor [10], the bi-factor model can shed light about APM scores being affected by other sources of variation in addition to *g*. Indeed, APM scores do not represent a perfect measure of *g* and that alternative tests (such as Arithmetic Applications from the Weschler Adult Intelligence Scale included in the Minnesota Study of Twins Reared Apart [11]) were more strongly loaded by *g* in some specific datasets [12]. Moreover, approximately 50% of the APM true variance could be related to *g*, with 10% belonging to specific factors, and as much as 25% related to test specific variance [12]. Confirmatory bi-factor models (i.e., BCFA) also presented a better fit to the data than the unidimensional model in alternative applications such as the Coloured Progressive Matrices test (an adaptation of the APM test to children from five to 11 years old; [13]).

Most recently, the presence of additional dimensions accounting for speed factors (as well as other effects such as item position) in APM scores [14] has been linked to specific learning types [15] as well as developmental differences [16]. In either case, such evidence reflects these factors possibly being of theoretical interest. Nevertheless, the presence and nature of these additional factors in APM scores is still a matter of contention.

1.2. Modern Approaches Towards Dimensionality Assessment

Most authors have generally based their decisions regarding the unidimensionality of the SPM scores either by applying eigenvalue-based dimensionality assessment methods (i.e., parallel analysis), by comparing fit statistics from CFA models (i.e., comparing the Comparative Fit Index) or by inspecting general factor reliability (i.e., Cronbach's α). Unfortunately, these three strategies have substantial shortcomings: Firstly, parallel analysis could hide relevant sources of variation while overestimating the presence of a single factor [17]. Also, its estimation is substantially affected by the response patterns when analysing tetrachoric and polychoric correlation matrices under limited sample size [18]. Secondly, CFA models could hide severe misspecification issues and result in biased parameter estimation [19,20]. Accordingly, CFA model-based reliability estimations could also be highly biased [21]. Thus, exploratory structures should be preferred in many cases [18,19]. We aim to resolve these issues by complementing these analyses with a new technique for dimensionality assessment (EGA) and the novel investigation of different exploratory factor models for the SPM-LS test.

1.2.1. Parallel Analysis

Parallel analysis is one of the main tools for dimensionality assessment [17,22,23]. Either when based on principal component or factor analysis solutions, parallel analysis has repeatedly been shown to optimally detect the true underlying unidimensionality in simulation studies [23–25]. However, parallel analysis is also fallible [18,23], with different conditions affecting each version of this procedure [17,22]. Principal component factor analysis is more reliable than the factor analysis alternative for structures with a small number of factors and binary data [17,22]. Unfortunately, it tends to wrongly suggest a single component to be retained if high factor correlations are present (as expected to occur in SPM-LS; [3]). On the other hand, factor analysis-based parallel analysis could be misleading if factors are not well defined (i.e., factor loadings < 0.40; [17]), which is indeed a plausible scenario for SPM-LS scores based on [12] depiction of APM variance partition. Additionally, either method presents difficulties in recovering the true dimensionality if samples < 500 are analysed (the size of [2] dataset; [17,26]). Finally, binary and categorical items presenting highly unbalanced categories (e.g., where the correct response represents 80–90% of the observed responses) could strongly affect parallel analysis performance [18,27,28].

1.2.2. Exploratory Graph Analysis

Exploratory Graph Analysis (EGA) is a statistical procedure that assesses latent dimensionality by exploring the unique relationships across pairs of variables (rather than the inter-item shared variance, as in common factor analysis; [29]). To do so, a sparse Gaussian Graphical Model is estimated (i.e., GGM) over the K precision matrix. K is the inverse of the inter-item variance-covariance matrix (i.e., $K = \Sigma^{-1}$; [30]) and it contains the partial correlations across pairs of observed variables. The sparse GMM is estimated by applying a penalization function (a common method is to select the GMM which minimises the extended Bayesian Information Criterion). After the GLASSO GMM is estimated, a walktrap clustering algorithm is applied to detect the optimal number of clusters in the network and to assign each item to a single dimension [21]. This algorithm, namely the combination of GLASSO GMM and walktrap clustering, has received the name of EGA. Although alternative versions of EGA exist, such as EGA with the triangulated maximally filtered graph approach (EGAtmfg), the former is preferred when high correlations between factors are expected (being the case for SPM-LS) [21].

EGA has been successfully applied to investigating the dimensionality of constructs such as personality [31], intelligence [32], and demonstrated to be as effective as parallel analysis when recovering true dimensionality under dichotomous data [17]. Nonetheless, EGA should be able to detect the number of underlying dimensions equal to or better than parallel analysis, even under suboptimal conditions (limited sample size; [17]). EGA is not presented as a substitute for techniques such as parallel analysis, but rather as a complementary tool to be studied in combination with them [17]. Accordingly, if parallel analysis results in indications of multidimensionality, researchers could benefit from exploring new techniques based on network analyses [30].

1.2.3. Exploratory Bi-factor Modelling

A review of the SPM literature has shown that two main factors models have been of interest: a unidimensional [2,4] and a multidimensional (bi-dimensional) solution [8]. Thus, it is legitimate to question to what extent specific sources of variance detected by parallel analysis or EGA could provide additional, meaningful information beyond g. In this sense, the bi-factor model should be the model to be evaluated [32,33]. The bi-factor model has been depicted as the best-suited model for assessing variance partition, to examine whether a structure is sufficiently unidimensional, and to measure the incremental value of potential specific factors [21,32,33]. When assessing estimated general factor strength, factor reliability should be compared using the omega hierarchical statistic (ω_H) [21,32]. Additionally, and to test the hypothesis of sufficient unidimensionality, the Explained Common Variance (i.e., ECV) and the Percentage of Uncontaminated Variances (PUC) should be compared altogether with ω_H for confirmatory models [34,35][1].

All model-based statistics are computed from a standardised factor analysis solution [32,36]. Therefore, it is necessary to ensure a proper estimation of the underlying bi-factor model in order to obtain unbiased reliability and ECV estimates. Given the difficulties for CFA models to recover complex structures (such as the bi-factor model) under realistic conditions (when cross-loadings are expected to occur; [19]), the bi-factor CFA models are often expected to produce biased parameter estimation [33]. In this context, exploratory alternatives such as EFA or Exploratory Structural Equation Modeling (i.e., ESEM) are becoming more and more widespread [37,38]. As these techniques offer model fit assessment while not imposing restrictions on the factor pattern matrix, they provide the modelling advantages of CFA while improving parameter estimation [18,39].

Exploratory bi-factor analysis (BEFA; Panel D, Figure 1) is a widely applied, compelling alternative to confirmatory bi-factor models [40]. The unique distinction between a BCFA and BEFA is that the latter allows the presence of cross-loadings for all specific factors [36] while maintaining the remaining characteristics (i.e., orthogonality between all factors). As each specific factor is still expected to be loaded by at least three indicators, variance partition, as well as the remaining BCFA characteristics, are present in a BEFA model [35]. However, how to approximate BEFA models is still a matter of debate. One of the most promising alternatives is via bi-factor target rotation, a technique applied in the BIFAD [10], the PEBI [41], or the SL-based iterative target rotation (SLi and SLiD algorithms; [36,38]).

In bi-factor target rotation, factor loadings to be minimised in the rotation procedure (i.e., items expected to have near-zero magnitude in the rotated loading matrix) are identified by giving them a zero value in the target matrix. As a convention, as general factor loadings are always freed (as each loading is expected to have a substantial load on this factor). The main issue then is to identify which loadings should be freed in the target rotation for the specific loadings. Conveniently, empirical cut-off points such as promin [42] or the procedure applied in SLiD algorithm [36] are able to select which loadings to be fixed based on each factor 's loadings distribution, and to prevent researchers

[1] Specific factor omega hierarchical and PUC are only computable for confirmatory solutions. Estimating such statistics in exploratory models would require researchers to decide which items or correlations are being considered by the specific factors.

from deciding on applying inappropriate fixed cut-off points (such as fixing all $\lambda < 0.20$; [36]). As an example, SLiD has been demonstrated to accurately recover bi-factor models in conditions under realistic conditions (i.e., cross-loadings or specific loadings of near-zero value), and to outperform more well-known methods such as the Schmid-Leiman orthogonalization, and the family of analytic rotations [43,44]. Promin-based algorithms (i.e., PEBI) has also been depicted as a compelling alternative and an improvement over alternative algorithms such as BIFAD [42]. Additionally, as the use of empirically defined target rotation is expected to improve parameter estimation, the estimation of general omega hierarchical, ECV and other model-based reliability estimates is also anticipated to be improved.

1.3. SPM-LS Dimensionality

SPM-LS dimensionality was evaluated by using a combination of parallel analysis, EFA and CFA results [2]. However, due to the limited sample size and the unbalanced responses patterns, parallel analysis results presented by the authors should be examined with caution. As the authors acknowledged, SPM-LS data presented some strong ceiling effects, when "10.4% of the sample had a perfect score of 12" [2] (p.114). This situation could have resulted in suboptimal performance of parallel analysis. In the results section, the authors declared that up to five factors should be retained via factor analysis parallel analysis. Additionally, and due to the large ratio of the first to second eigenvalue (5.92 to 0.97), evidence of a robust general factor was said to be found [2]. However, as factor analysis parallel analysis could be more unreliable than its principal-component alternative for the study at hand (due to limited sample size and the binary nature of the data), the results of both techniques should have been taken into consideration (e.g., when computing ratios of eigenvalues).

The authors additionally reported that no evidence of relevant specific factors was identified, as factor pattern loadings on unreported solutions including two to five factors were not in line with any theoretical expectation (i.e., "were uninterpretable"; [2], p. 112). However, the authors did not report the structures tested, or if models combining general and specific sources of variation (i.e., bi-factor) were estimated. Lastly, as global fit indexes suggested an adequate fit for the unidimensional model (i.e., even though RMSEA was as high as 0.079) and the general factor was considered as reliable ($\omega_H = 0.86$), the authors concluded that the SPM-LS scores could be considered essentially unidimensional [2] (p.112). In this investigation, this claim will be revisited by a more nuanced inspection of SPM-LS scores by applying traditional methods (exploratory and confirmatory unidimensional and bi-dimensional factor models) as well as two recently developed methods for assessing and validating multidimensional scales (EGA and bi-factor exploratory modelling).

2. Materials and Methods

2.1. Instrument and Data

The SPM-LS scores are those made publicly available by [2] for this special edition. In detail, the sample is composed of the answers of 499 undergraduate students who responded to the SPM-LS. The SPM-LS consists of the last 12 matrices the Standard Progressive Matrices [1] (i.e., those of greatest difficulty). Noteworthy, even though these items could be considered as polytomous, and essential information could be retrieved if they were treated as such [2], it is common to score them as dichotomous items: either a respondent identified the correct answer or not according to the item key provided by the authors. Accordingly, the tetrachoric correlation matrix was here studied. In this application, respondents had no time limit to complete the 12 items and were encouraged to respond to each item. Accordingly, no missing data were observed.

2.2. Statistical Analysis Plan

The following analysis will be performed to inspect the factor structure of the SPM-LS: Firstly, the dimensionality of the SPM-LS will be assessed applying both, principal component and factor analysis parallel analysis. Secondly, these results will be contrasted with those of EGA. If the SPM-LS is regarded as multidimensional, the hypothesis of essential unidimensionality will be tested by inspecting a series of unidimensional, exploratory and confirmatory bi-dimensional and bi-factor models (Figure 1). These models would be compared in terms of model fit, factor pattern results, ω_H and ECV, and PUC values (when possible). To estimate BEFA models, a bi-factor target rotation would be defined from bi-dimensional EFA solution, using the empirical cut-off point definition algorithm included in SLiD [36] and the promin cut-off estimation [42].

Most analyses were conducted in R 3.5.2. [45] in a reproducible manner using the rmarkdown [46] and the papaja [47] packages. The correlation matrix was obtained using the *cor_auto ()* function in the qgraph package [48], which provided similar results to the *tetrachoric ()* function from the psych package [49]. Principal component and factor analysis were conducted using the *fa.parallel ()* function in the psych package [49]. EGA was applied using the EGA package [50]. EFA and CFA models were computed using the lavaan package [51]. Cronbach's α and omega estimates were computed from the *reliability ()* function from the semTools package [52] following current recommendations on the field [53]. EFA models were rotated using oblique target rotation using the gradient projection algorithm included in the GPArotation package [54]. Bi-factor target was defined using the promin rotation [42] and the algorithm included in the SLiD [36]. The bi-dimensional EFA model was computed using minimum residual as the extraction method and target rotation towards the expected EGA solution. ESEM models for estimating bi-dimensional EFA and bi-factor EFA models with a free residual correlation were fitted in Mplus 7.3. Scripts for reproducing all analyses (i.e., main text, Appendices A and B results) can be found as Supplementary Data.

3. Results

3.1. Descriptive Analysis

A characteristic of the SPM-LS is that the chosen items represent the most difficult items from the SPM. However, the proportion of correct responses did not monotonically decrease as a function of item position (Figure 2), as it could be somewhat expected. The first six items (SMP1 to SMP6) had high correct proportions of correct responses ($0.76 < p_{correct} < 0.91$; where $p_{correct}$ is the observed proportion of correct answers) and were identified to present similar rates of unbalanced response patterns. On the other hand, the last three less than half of the responses collected were correct items (SPM10: $p_{correct} = 0.39$; SPM11: $p_{correct} = 0.36$ and SPM12: $p_{correct} = 0.32$). As said before, these unbalanced response patterns could lead to significant estimation errors in the tetrachoric correlation estimation.

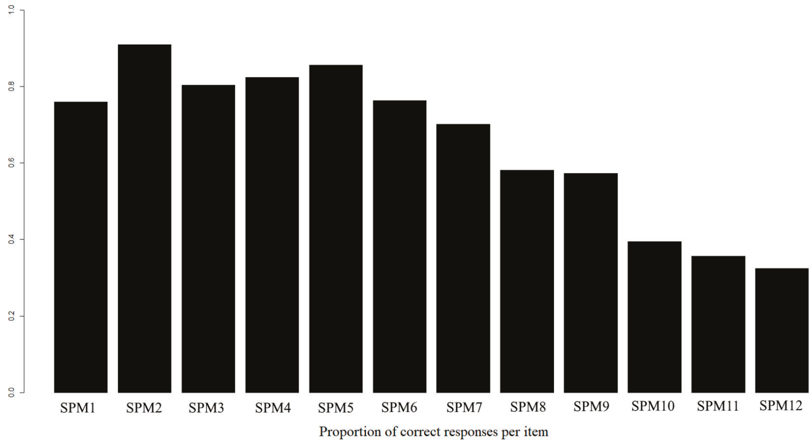
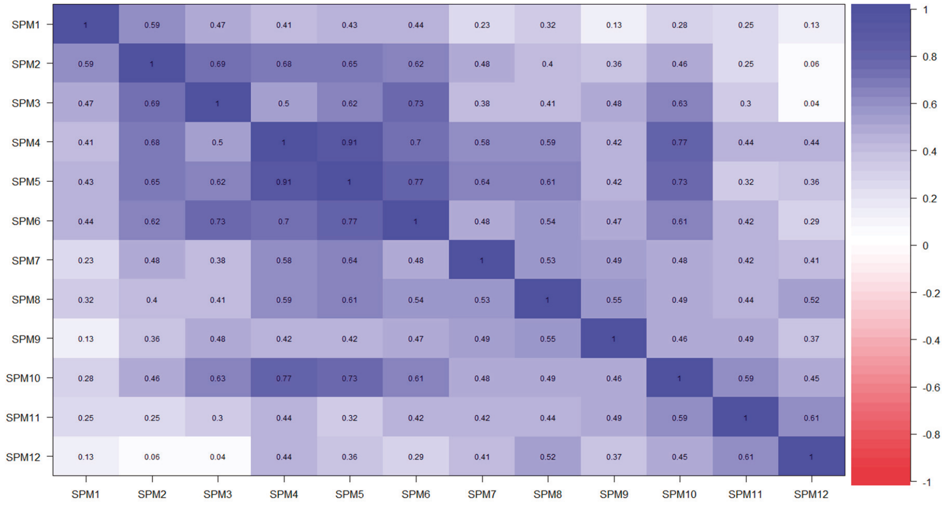


Figure 2. Proportion of correct responses as a function of item location in the SPM-LS.

A visual inspection of the tetrachoric correlation matrix (Figure 3) revealed an unusually high correlation between items ($r_{\text{SPM4}-\text{SPM15}} = 0.91$), which was substantially larger than the ensuing correlation in terms of magnitude ($r_{\text{SPM5}-\text{SPM16}} = 0.77$). In detail, 79.8% of individuals who correctly responded SPM4, also were correct for SPM5. Moreover, 11.8% of respondents who failed SPM4, also failed SPM5. Thus, there was only 8.4% of respondents who failed/gave a correct answer or gave a correct answer/failed SPM4-SPM5, respectively. A visual inspection of the tetrachoric correlation heatmap revealed two distinct blocks of inter-item correlations: The first one between items SMP1 to SPM6, and the second one between items SPM7 to SMP11. Therefore, Figure 3 is indicative of two distinct sources of multidimensionality. Due to the limited sample size, and the highly unbalanced response patterns for items such as SPM2, SPM11, and SPM12, it is noteworthy that the tetrachoric correlations between these items could be affected by significant estimation errors.



	SPM1	SPM2	SPM3	SPM4	SPM5	SPM6	SPM7	SPM8	SPM9	SPM10	SPM11	SPM12
SPM1	1	0.59	0.47	0.41	0.43	0.44	0.23	0.32	0.13	0.28	0.25	0.13
SPM2	0.59	1	0.69	0.68	0.65	0.62	0.48	0.4	0.36	0.46	0.25	0.06
SPM3	0.47	0.69	1	0.5	0.62	0.73	0.38	0.41	0.48	0.63	0.3	0.04
SPM4	0.41	0.68	0.5	1	0.91	0.7	0.58	0.59	0.42	0.77	0.44	0.44
SPM5	0.43	0.65	0.62	0.91	1	0.77	0.64	0.61	0.42	0.73	0.32	0.36
SPM6	0.44	0.62	0.73	0.7	0.77	1	0.48	0.54	0.47	0.61	0.42	0.29
SPM7	0.23	0.48	0.38	0.58	0.64	0.48	1	0.53	0.49	0.48	0.42	0.41
SPM8	0.32	0.4	0.41	0.59	0.61	0.54	0.53	1	0.55	0.49	0.44	0.52
SPM9	0.13	0.36	0.48	0.42	0.42	0.47	0.49	0.55	1	0.46	0.49	0.37
SPM10	0.28	0.46	0.63	0.77	0.73	0.61	0.48	0.49	0.46	1	0.59	0.45
SPM11	0.25	0.25	0.3	0.44	0.32	0.42	0.42	0.44	0.49	0.59	1	0.61
SPM12	0.13	0.06	0.04	0.44	0.36	0.29	0.41	0.52	0.37	0.45	0.61	1

Figure 3. Heatmap of SPM-LS items tetrachoric correlation.

3.2. Dimensionality Assessment.

We exactly replicated the results provided by [2] when computing parallel analysis over the tetrachoric correlation matrix (using maximum likelihood)[2] (Left panel, Figure 4; also Figure 1 in [2]). The number of factors to be retained was 5, with eigenvalues of 5.92, 0.93, 0.36, 0.18, and 0.10 (simulated eigenvalues of.52, 0.21. 0.16, 0.12, 0.07). The number of components to be retained was 2, with eigenvalues as of 6.36 and 1.60 (simulated eigenvalues of 1.26 and 1.20). Noteworthy, it was observed that the authors conducted this analysis over the tetrachoric correlation matrix, obtaining the eigenvalues to be compared against those extracted by generating random normal data. However, this strategy is considered highly inadequate [18]. A better strategy when analyzing tetrachoric correlations is to obtain the random eigenvalues by resampling from the observed data. Accordingly, we repeated the analysis with this specification (Right panel, Figure 4). Factor and principal component factor analysis suggested to retain two and three factors/components, respectively: factor analysis parallel analysis showed eigenvalues of 3.43, 0.73 and 0.33 (with resampled eigenvalues of 0.54, 0.20 and 0.15) while principal components PA resulted in eigenvalues of 4.09, 1.51 for the original components (with resampled components of 1.26 and 1.19).

[2] Using other extraction methods (i.e., ordinary least squares) led to similar conclusions regarding the underlying dimensionality, but for weighted and generalized least squares, which suggested to retain three factors and two components.

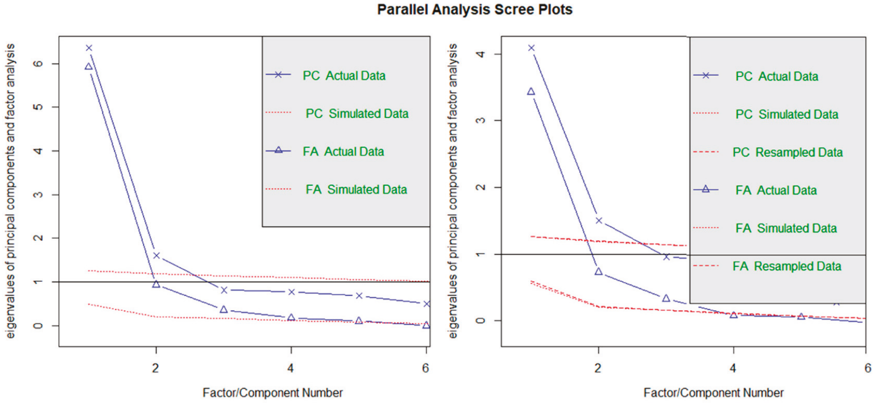


Figure 4. Parallel analysis results: (**a**) Original Principal component and parallel factor analysis with eigenvalue simulated from random normal data; (**b**) Principal component and parallel factor analysis correct eigenvalues obtained from resampling from original data.

Nevertheless, both parallel analysis techniques are suggesting the SPM-LS be multidimensional. The discrepancy between both methods (suggestions of three factors vs two components to be retained) could be due factor analysis-based parallel analysis being more affected by the limited sample size analysed. EGA agreed with principal component parallel analysis and identified two underlying dimensions (Figure 5), one composed of items one to six and the other of items seven to twelve. Moreover, EGA results confirmed that the highest observed partial correlation was observed for the pair SPM4–SPM5. This partial correlation indicates that, after controlling for all the other variables, these items were strongly conditionally dependent.

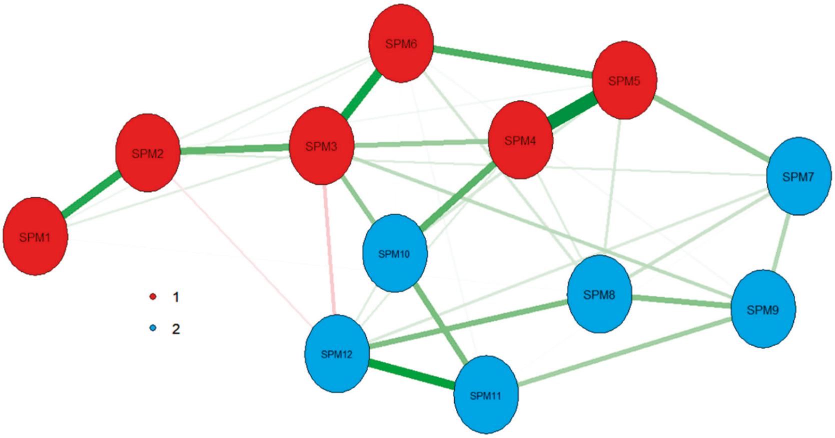


Figure 5. Exploratory Graph Analysis of SPM-LS data. Dimensions and items associated are presented in different colours. Positive partial correlations are depicted in green, with negative partial correlations presented in red. The size of the lines indicates the size of the partial correlations.

Therefore, and after inspecting the tetrachoric correlation matrix and observing the dependence between SPM4–SPM5 items, it was decided to reanalyse SPM-LS dimensionality after aggregating these items. Item parcelling (i.e., aggregating items) have been shown as a valid alternative to deal with residual item covariances [55]. Both techniques of parallel analysis agreed in this re-analysis that two factors should be retained. EGA also resulted in two factors being identified, with a similar distribution than in Figure 5. Therefore, robust evidence from both, parallel analysis and EGA, supported the hypothesis of SPM-LS being bi-dimensional (either when treating the original set of items, or the reduced version combining items SPM4 and SPM5). Analysis details and results of this analysis are presented in Appendix A.

3.3. Factor Modelling

The standardised factor solutions for all estimated models are shown in Table 1. Likewise, the fit indices for all estimated models are presented in Table 2. For the sake of comparison, similar models not estimating the residual correlation between SPM4–SPM5 were also computed. Standardised factor loadings and model fit indices of these models without including this residual correlation are presented in Appendix B.

Table 1. Standardized factor loadings for all model tested.

	Unidim.	**Unidim.M.**	**BID.EFA**		**BID.CFA**		**BEFA**		**BCFA**	
Item	G	G	S1	S2	S1	S2	G	S1	G	S
SPM1	**0.47**	**0.48**	**0.59**	−0.08	**0.50**	0.00	**0.54**	−0.12	**0.50**	0.00
SPM2	**0.72**	**0.74**	**0.93**	−0.15	**0.76**	0.00	**0.82**	−0.22	**0.77**	0.00
SPM3	**0.72**	**0.73**	**0.88**	−0.09	**0.76**	0.00	**0.81**	−0.16	**0.76**	0.00
SPM4	**0.92**	**0.84**	**0.55**	**0.41**	**0.89**	0.00	**0.81**	0.25	**0.88**	0.00
SPM5	**0.94**	**0.87**	**0.67**	**0.30**	**0.91**	0.00	**0.86**	0.15	**0.91**	0.00
SPM6	**0.81**	**0.83**	**0.75**	0.17	**0.85**	0.00	**0.85**	0.05	**0.85**	0.00
SPM7	**0.66**	**0.67**	**0.30**	**0.47**	0.00	**0.71**	**0.60**	**0.33**	**0.62**	**0.30**
SPM8	**0.70**	**0.71**	0.23	**0.58**	0.00	**0.75**	**0.61**	**0.42**	**0.62**	**0.40**
SPM9	**0.61**	**0.61**	0.20	**0.50**	0.00	**0.65**	**0.53**	**0.36**	**0.52**	**0.39**
SPM10	**0.79**	**0.80**	**0.43**	**0.48**	0.00	**0.85**	**0.74**	**0.32**	**0.76**	0.27
SPM11	**0.62**	**0.63**	−0.04	**0.75**	0.00	**0.66**	**0.44**	**0.58**	**0.44**	**0.63**
SPM12	**0.53**	**0.54**	**−0.38**	**1.00**	0.00	**0.57**	0.28	**0.80**	**0.31**	**0.73**
φ	-	-	0.56		0.82		0.00		0.00	
SPM4-SPM5	-	0.69	0.70		0.56		0.70		0.57	

[1] Unidim = Unidimensional model. Unidim.M. = Unidimensional model with SPM4-SPM5 residual correlation estimated. BID.EFA = Bi-dimensional exploratory factor analysis. BID.CFA = Bi-dimensional confirmatory factor analysis. BEFA = Bi-factor exploratory factor analysis. BCFA = Bi-factor confirmatory factor analysis. All loadings over 0.30 are presented bolded. φ = Inter-factor correlation. SPM4–SPM5 = Residual covariance between SPM4-SPM5 items. G = General Factor. S1= First specific factor. S2 = Second specific factor. Factor loadings with values > 0.30 appear bolded.

Table 2. Model fit indices for all tested models.

	Np	**df**	X^2	p	**CFI**	**TLI**	**RMSEA**	**SRMR**
Unidim.	24	54	221.75	0.00	0.95	0.93	0.08 (0.07–0.09)	0.11
Unidim.M.	25	53	205.88	0.00	0.95	0.94	0.08 (0.07–0.08)	0.11
BID.EFA/BEFA.	36	42	**80.50**	**0.00**	**0.99**	**0.98**	**0.04 (0.03–0.06)**	**0.06**
BID.CFA	26	52	160.69	0.00	0.96	0.96	0.07 (0.05–0.08)	0.09
BCFA	31	47	113.72	0.00	0.98	0.97	0.05 (0.04, 0.07)	0.07

[1] Unidim = Unidimensional model. Unidim.M. = Unidimensional model with SPM4–SPM5 residual correlation estimated. BID.EFA = Bi-dimensional exploratory factor analysis. BID.CFA = Bi-dimensional confirmatory factor analysis. BEFA = Bi-factor exploratory factor analysis. BCFA = Bi-factor confirmatory factor analysis. Np = Estimated number of parameters. Df = degrees of freedom. ^{2} = Chi-square statistic. P = p-value associated with ^{2} test of fit. CFI = Comparative fit index. TLI = Tucker-Lewis index. RMSEA = Root Mean Square Error of Approximation (with 95% confidence interval in parenthesis). SRMS = Standardized Root Mean Square Residual. Best fit indices presented bolded and underlined. Model fit indices for the best fitting model appear bolded.

3.3.1. Unidimensional Model

We first replicated the original results with regards to the CFA unidimensional model [2]. We found the same model fit indices (CFI = 0.95, TLI = 0.93, RMSEA = 0.08, SRMS = 0.11). Cronbach's $\alpha = 0.92$ and $\omega_{HG} = 0.83$ also matched those reported. For this model, the high RMSEA and SRMR values suggest questionable fit. Estimating the correlation between SPM4–SPM5 resulted in improved model fit (CFI = 0.95, TLI = 0.94, RMSEA = 0.08, SRMS = 0.11). As expected, the SPM4–SPMP5 correlation was high and positive ($\psi = 0.69$). Accordingly, the remaining presented models will include the estimation of the residual correlation between both items. Additionally, this unidimensional model showed adequate reliability (Cronbach's α = 0.92; ω_{HG} = 0.86).

3.3.2. Bi-Dimensional Model

Two bi-dimensional structures were computed. Firstly, an exploratory bi-dimensional model was fitted in order to understand if EFA results supported the idea of a bi-dimensional SPM-LS structure. Secondly, such an EFA structure was tested as a confirmatory model to understand the role of potential cross-loadings present on the data. EFA model fit indexes revealed that this structure provided an excellent fit to the data (CFI = 0.99, TLI = 0.98, RMSEA = 0.04, SRMS = 0.06), improving model fit with respect to the unidimensional case. Additionally, a lower inter-factor correlation of (φ ≈ 0.56) was obtained[3]. The SPM4–SPM5 correlation of this residual correlation ($\psi = 0.70$) was similar to the one observed in the unidimensional model.

The confirmatory bi-dimensional (CFI = 0.96, TLI = 0.96, RMSEA = 0.06, SRMS = 0.09) presented a better model fit than the unidimensional model, but worse than its exploratory counterpart. Fixing all cross-loadings to zero led to observe a larger factor correlation (φ = 0.82), larger SPM4–SPM5 loadings ($\lambda_{SPM4} = 0.89, \lambda_{SPM5} = 0.91$), and a diminished residual correlation between them ($\psi = 0.56$). In this case, both factors were considered as reliable if measured by Cronbach's α standards (factor 1 = 0.91, factor 2 = 0.85), and close to acceptable reliability when inspecting ω_{HS} (factor 1 = 0.75 factor 2 = 0.70). In conclusion, a bi-dimensional model (either by EFA/CFA based) improved model fit over the unidimensional structure. As indicated by the substantial inter-factor correlation observed in all models, a general factor could play a substantial role in SPM-LS structure. This hypothesis will be explored next via bi-factor modelling.

3.3.3. Bi-Factor Model

Two bi-factor models were tested: a BEFA model fitted using bi-factor target rotation and a BCFA model restricting cross-loadings to zero. Either using the algorithm included in SLiD [36] or a promin-based cut-off [42] resulted in items SPM7 to SPM12 being freed in the specific factor. Noteworthy, as rotation does not affect model fit [29], fit indices for this model were those of the exploratory bi-dimensional structure. The BEFA model (Table 1) presented three main characteristics: (a) The rotation procedure recovered orthogonal factors (even if oblique target rotation was applied), which aligns with the expectations of the bi-factor model; (b) Although the general factor was well-defined (all loadings over $\lambda_G > 0.30$), SPM11 and SPM12 presented higher loadings on the specific factor ($\lambda_{SSPM11} = 0.58, \lambda_{SSPM12} = 0.80$) than in the general factor ($\lambda_{gSPM11} = 0.44, \lambda_{GSPM12} = 0.29$); (c) the residual correlation between SPM4 and SPM5 was similar to the one observed for the unidimensional model ($\psi = 0.70$). With regards to BEFA general factor reliability, it was considered as adequate (ω_{HG} = 0.80; ECV = 0.74).

The BCFA model showed the best fit indexes from all confirmatory models (Table 2; CFI = 0.98, TLI = 0.97, RMSEA = 0.05, SRMS = 0.07). Both factors were well-defined (all loadings $\lambda > 0.30$) with SPM4–SPM5 general loadings being stronger than in the BEFA model (as they were inflated due their cross-loadings being fixed to zero). SPM4–SPM5 residual correlation was similar to the one observed in the confirmatory bi-dimensional model ($\psi = 0.57$). Overall, general factor reliability was also adequate (ω_{HG} = 0.75; ECV = 0.80). Additionally, the associated PUC was $(132 - 42)/132 = 0.68$. Under the presence of $PUC < 0.80$, researchers are recommended that $\omega_H > 0.70$ and $ECV > 0.60$

[3] Using alternative oblique rotations (i.e., oblimin, promax, geomin) resulted in factor structures with a similar distribution of loadings and size. Main differences were small in magnitude, and mostly affected the inter-factor correlation size.

be used as benchmarks for considering essential unidimensionality [34]. Therefore, while the BCFA provided an adequate approximation towards SPM-LS multidimensionality, the presence of a strong, reliable general factor also favours that SPM-LS scores be considered as essentially unidimensional. Lastly, the specific factor reliability (ω_{HS} = 0.31) was in the range of values commonly observed on bi-factor modelling [32,33].

4. Discussion

The SPM-LS (Standard Progressive Matrices–Last Series) has been recently proposed as an improved short version of the SPM test [2]. The SPM-LS was treated as an essentially unidimensional measure of *g*, with better psychometric properties than alternative tests such as the Advanced Progressive Matrices test (i.e., APM). On these grounds, [2] proceeded to fit a series of IRT models to study the benefits of studying the nominal responses in the test, acknowledging that mixed results from EFA and CFA results could suggest SPM-LS not being a strictly unidimensional measure. The authors further recommended investigators to conduct additional research on this matter. We aimed to shed light on SPM-LS dimensionality using improving the dimensionality techniques applied (comparing parallel analysis with exploratory graphic analysis results) and by providing a thoughtful exploration of unidimensional, bi-dimensional and bi-factor SPM-LS structures.

The main result of this study is that SPM-LS can be considered as essentially unidimensional measurement of intelligence if appropriately treated. Reliability and unidimensionality indices obtained from a bi-dimensional bi-factor model provided strong evidence of this conclusion. Notwithstanding the evidence of essential unidimensionality, it is also true that a non-ignorable, nuisance factor associated with the last six indicators of the SPM-LS was systematically found, either when applying parallel analysis, EGA, or factor modelling. An additional residual covariation between SPM4–SPM5 was also observed. This circumstance that should be discussed in more detail: Firstly, such a high residual correlation between both items might be due to significant estimation error in the tetrachoric matrix, altogether with the limited sample size. If so, future research employing different, larger samples should be able to identify a substantially smaller covariation between these items. Secondly, the relationship between SPM4 and SPM5 in terms of content and rules used for resolving these items should be inspected in further detail in order to decide if the information provided by both items is truly distinct or redundant.

This study evidence dimensionality assessment is a complex task which often requires convergent evidence from different sources and statistical techniques (as suggested in the case of parallel analysis and EGA; [17]). Moreover, being overconfident about model fit indices could be misleading when selecting an appropriate solution. Model fit should always be complemented with alternative indices (such as ω_H, ECV or PUC) when possible [34]. Lastly, caution should be exercised when interpreting high inter-factor correlations in confirmatory models as evidence of unidimensionality, as these correlations could be inflated if relevant cross-loadings are being omitted. As an example, the inter-factor correlation was substantially larger for the bi-dimensional confirmatory structure that for its exploratory counterpart. To avoid such situations, we recommend researchers to confront results from both exploratory and confirmatory versions of the models to be investigated. If relevant cross-loadings to be potentially fixed are identified, we agree with previous authors that exploratory models should be prioritized [19,20].

Lastly, the result of applying bi-factor modelling was clear: We found evidence of a robust and reliable *g* factor (which resulted in our conclusion of SPM-LS scores being essentially unidimensional by current benchmarks [34]), plus an additional nuisance factor related with the last six items. While the interpretation of this latter factor could be somewhat controversial, it cannot be associated with a speed factor as in previous applications of similar tests [7,56] (as respondents had no time limit to reply to the matrices). An alternative explication is that such a factor would be related to guessing strategy or a difficulty component. Noteworthy, the first six items were (almost uniformly) correctly responded (with a proportion of correct responses near to 0.80), with the last six items presented a decreasing proportion of right answered (as evidenced in Figure 2). Under these conditions, it is known that parallel analysis is set to fail and that exploratory factor analysis under tetrachoric correlations could result in reflecting a difficulty factor [57,58]. Alternatively, the idea of guessing strategies being a relevant aspect of SPM-LS data was strongly supported by the original authors [2], as they showed that a three-parameter IRT model (incorporating a pseudo-guessing parameter) fitted the data better than alternative models. In this sense, and as pointed out by a reviewer, statistical artefacts of similar nature could be observed when applying factor analysis to a tetrachoric correlation matrix obtained from data generated from a three-parameter IRT model. Therefore, additional research on this matter should be granted in future SPM-LS applications. Thus, evidence suggests that guessing could play a substantive role with regards to general intelligence estimation. Even though we expanded these findings by identifying that guessing could also affect dimensionality assessment, future research should focus on re-assessing SPM-LS dimensionality under the assumption of data being generated from the three-parameter nested logistic model, as it has been shown to improve the effectiveness of parallel analysis [58]. Lastly, specific item position and item difficulty effects should aim to be separately studied (as they are confounded in the current SPM-LS form). Additionally, structural models aimed to measure each specific effect should also be encouraged to be applied [14].

Overall, the consequences of the presented findings are two-folded: firstly, even though researchers could treat SPM-LS as essentially unidimensional, this does not preclude them to not use the better measurement model (i.e., the bi-factor form) in their statistical analyses, especially if included within an SEM framework. Failing to take the influence of the second factor into account could lead to inflating or deflated regression coefficient and other types of measurement error propagation [39]. As an example, in our results, the variance explained by the second factor is of 0.17. If we assume a criterion Y, measured with reliability of one and a perfect positive relationship with the nuisance factor, the expected value for the estimated correlation between our nuisance factor and Y would be estimated as 0.41 (considering the attenuation by reliability described in [59]). Even though such distorting effect represents a worst-case scenario, where expected attenuation effects are anticipated to be smaller (as either criterion reliability or true relationship between criterion or specific factor would be not perfect), they should not be disregarded as negligible [59].

An attenuation of this magnitude could impact the evaluation of SPM-LS scores criterion and incremental validity (the expected increment of the determination coefficient might range from zero to 0.17). Note that our analysis identifies a source of performance variance. The effects might be even more substantial for a group with larger variance in the secondary factor. Consequently, despite the essential unidimensionality of the measure, the consequences of taking or not this second factor into account must be weighted in future research endeavours, including additional intelligence and ability measures.

Secondly, and from a theoretical point of view, researchers should not automatically disregard such secondary factors, as they could be tied to relevant individual differences of the test-takers [15,16]. On the contrary, more research is needed for us to have a better understating of the nature of this nuisance factor, and the extent that it could represent valuable information of the examinees.

5. Conclusions

The SPM-LS has been suggested to be a valid, reliable alternative version of the Standard Progressive Matrices test, presenting superior psychometric properties to alternatives such as the Advanced Progressive Matrices test. In this research, we provided a detailed study of the essential unidimensionality claimed by the original authors by utilising applying modern dimensionality techniques and bi-factor modelling. Our results suggest that, if appropriately treated, SPM-LS scores can be considered as such. Nevertheless, an additional factor relevant to the last six items was identified. Additionally, we recommend evaluating further the presence of this factor in additional, larger sample sizes presenting more balanced responses to the SPM-LS test.

Supplementary Materials: The following are available online at http://www.mdpi.com/2079-3200/7/3/14/s1.

Author Contributions: E.G.-G.: Conceptualized the study; E.G.G., F.J.A., & L.E.G.: Established the overarching framework; E.G.-G., F.J.A., & L.E.G.: conducted the statistical analysis; E.G.-G.: Wrote the manuscript; F.J.A., & L.E.G.: Editing & approved the manuscript.

Funding: This research was by Ministry of Education, Culture and Sports (Spain) under Grant FPU15/03246 and by Ministerio de Economia y Competitividad (Spain) and Fondo Social Europeo under Grant PSI2013-44300-P and Grant PSI2017-85022-P.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

In this Appendix A, the SPM-LS dimensionality will be re-analysed by including a parcel created by aggerating SPM4-SPM5 items. This decision was taken based on the high dependence observed between items SPM4–SPM5 (i.e., tetrachoric correlation of 0.91; high partial correlation detected in EGA) Thus, we will follow the same steps performed in the primary analysis. Firstly, we reproduce the tetrachoric-polychoric correlation analysed in these analyses. As expected, most correlations between items and the combined item (i.e., SPM4-5) were like the original (Table A1).

Table A1. Tetrachoric/polychoric correlation matrix with SPM4 and SPM5 combined.

	1	2	3	4–5	6	7	8	9	10	11	12
SPM1	1										
SPM2	0.59	1									
SPM3	0.47	0.69	1								
SPM4-5	0.40	0.67	0.54	1							
SPM6	0.44	0.62	0.73	0.72	1						
SPM7	0.23	0.48	0.38	0.62	0.48	1					
SPM8	0.32	0.40	0.41	0.60	0.51	0.53	1				
SPM9	0.13	0.36	0.48	0.41	0.47	0.49	0.55	1			
SPM10	0.28	0.46	0.63	0.77	0.61	0.48	0.49	0.46	1		
SPM11	0.25	0.25	0.31	0.42	0.42	0.42	0.44	0.49	0.59	1	
SPM12	0.13	0.06	0.04	0.43	0.29	0.41	0.52	0.37	0.45	0.61	1

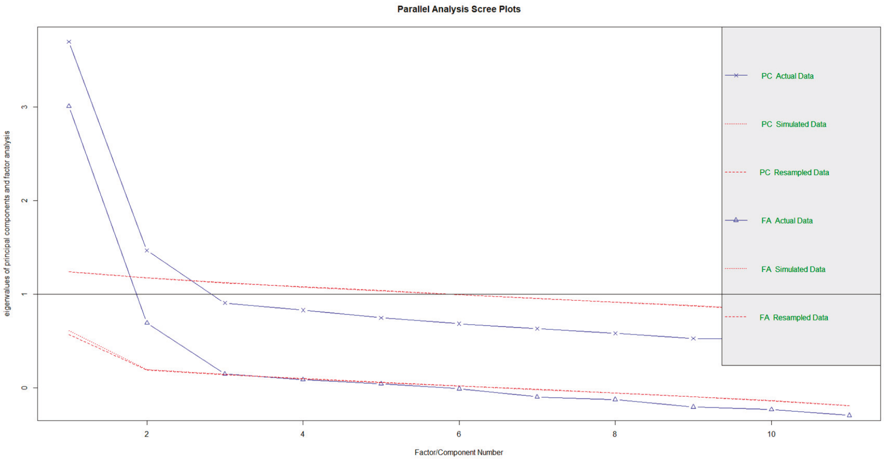


Figure A1. Principal component and parallel factor analysis with eigenvalue obtained from resampling from original data using a parcel for SPM4 and SPM5 items.

We performed principal components, and factor analysis parallel analysis with eigenvalues resampled from the original data over this correlation matrices. Both techniques agreed to indicate that the structure was bi-dimensional (Figure A2). The value of the original components was 3.70 and 1.47 (with resampled components of 1.24 and 1.17), and the value of the original factor was 3.01 and 0.69 (with resampled eigenvalues of 0.64 and 0.19).

EGA agreed with parallel analysis results and concluded that two dimensions are underlying the SPM-LS scores if SPM4 and SPM5 items were combined. Thus, there was robust evidence of the bi-dimensional nature of the data after controlling for the dependency between SPM4 and SPM5 items.

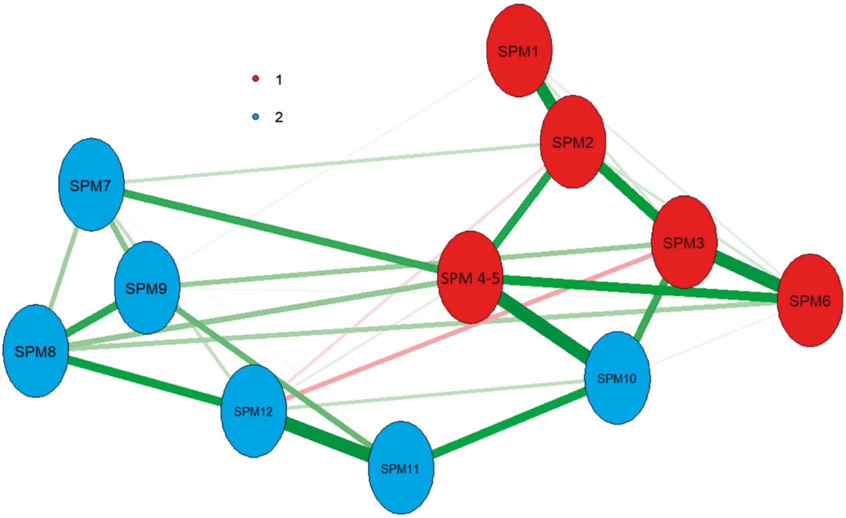


Figure A2. Exploratory Graph Analysis of SPM-LS data with SPM4 and SPM5 item combined. Dimensions and items associated are presented in different colours. Positive partial correlations are depicted in green, with negative partial correlations presented in red. The size of the lines indicates the size of the partial correlations. SPM4 = SPM4-5 item.

Table A2. Standardised factor loadings for all model tested.

	Unidim.	**BID.EFA**		**BID.CFA**		**BEFA**		**BCFA**	
Item	G	S1	S2	S1	S2	G	S1	G	S
SPM1	**0.47**	**0.58**	−0.04	**0.50**	0.00	**0.55**	−0.10	**0.51**	0.00
SPM2	**0.72**	**0.90**	−0.10	**0.76**	0.00	**0.84**	−0.17	**0.77**	0.00
SPM3	**0.74**	**0.87**	−0.04	**0.77**	0.00	**0.84**	−0.13	**0.79**	0.00
SPM4-5	**0.85**	**0.55**	**0.43**	**0.90**	0.00	**0.82**	0.28	**0.90**	0.00
SPM6	**0.82**	**0.70**	0.22	**0.86**	0.00	**0.84**	0.10	**0.85**	0.00
SPM7	**0.67**	0.26	**0.51**	0.00	**0.70**	**0.57**	**0.36**	**0.60**	**0.31**
SPM8	**0.71**	0.20	**0.60**	0.00	**0.74**	**0.58**	**0.45**	**0.61**	**0.41**
SPM9	**0.62**	0.19	**0.52**	0.00	**0.65**	**0.52**	**0.38**	**0.52**	**0.40**
SPM10	**0.80**	**0.40**	**0.51**	0.00	**0.84**	**0.72**	**0.35**	**0.75**	0.29
SPM11	**0.64**	−0.04	**0.75**	0.00	**0.67**	**0.43**	**0.59**	**0.45**	**0.61**
SPM12	**0.54**	**−0.39**	**1.00**	0.00	**0.57**	0.23	**0.82**	0.29	**0.76**
φ	**-**	**0.54**		**0.81**		0.00		0.00	

[1] Unidim = Unidimensional model. Unidim.M. = Unidimensional model with SPM4-SPM5 residual correlation estimated. BID.EFA = Bi-dimensional exploratory factor analysis. BID.CFA = Bi-dimensional confirmatory factor analysis. BEFA = Bi-factor exploratory factor analysis. BCFA = Bi-factor confirmatory factor analysis. All loadings over 0.30 are presented bolded. φ = Inter-factor correlation. SPM4-SPM5 = Residual covariance between SPM4-SPM5 items. G = General Factor. S1= First specific factor. S2 = Second specific factor. Factor loadings with values > 0.30 appear bolded.

Table A3. Model fit indices for all tested models.

	Np	df	X^2	p	CFI	TLI	RMSEA	SRMR
Unidim.	23	44	192.04	0.00	0.93	0.91	0.08 (0.07–0.09)	0.11
BID.EFA/BEFA.	33	34	68.45	0.00	**0.98**	**0.97**	**0.05 (0.03–0.06)**	**0.05**
BID.CFA	24	43	145.71	0.00	0.95	0.94	0.07 (0.06–0.08)	0.09
BCFA	29	38	110.79	0.00	0.97	0.96	0.06 (0.04–0.07)	0.07

[1] Unidim = Unidimensional model. Unidim.M. = Unidimensional model with SPM4-SPM5 residual correlation estimated. BID.EFA = Bi-dimensional exploratory factor analysis. BID.CFA = Bi-dimensional confirmatory factor analysis. BEFA = Bi-factor exploratory factor analysis. BCFA = Bi-factor confirmatory factor analysis. Np = Estimated number of parameters. Df = degrees of freedom. ^{2} = Chi-square statistic. P = p-value associated with ^{2} test of fit. CFI = Comparative fit index. TLI= Tucker-Lewis index. RMSEA = Root Mean Square Error of Approximation (with 95% confidence interval in parenthesis). SRMS = Standardized Root Mean Square Residual. Best fit indices presented bolded and underlined. Model fit indices for the best fitting model appear bolded.

Lastly, and in the case to be of interest, standardised factor loadings and model fit indices are provided. Noteworthy, results were similar to other models presented in this article but provided a sustainably worse fit to the data. In the exploratory models, SPM4-5 showed lower factor loadings in the S1 (model BID.EFA) or G (model BEFA), and higher cross-loadings on the alternative factors. In the confirmatory models, SPM4-5 loadings were also closer to 0.90 than in the main text results. Overall, resulting structures were mostly similar to those analysed in the result section of the article.

Appendix B

In Appendix B, standardised factor loadings (Table A4) and model fit indices (Table A5) are provided for models without the residual correlation SPM4-SPM5.

Table A4. Standardised factor loadings for all model tested.

	Unidim.	BID.EFA		BID.CFA		BEFA		BCFA	
Item	G	S1	S2	S1	S2	G	S1	G	S
SPM1	**0.47**	**0.59**	−0.09	**0.50**	0.00	**0.53**	−0.13	**0.50**	0.00
SPM2	**0.72**	**0.93**	−0.18	**0.75**	0.00	**0.81**	−0.23	**0.76**	0.00
SPM3	**0.72**	**0.87**	−0.10	**0.75**	0.00	**0.80**	−0.16	**0.76**	0.00
SPM4	**0.92**	**0.65**	**0.38**	**0.94**	0.00	**0.90**	0.22	**0.93**	0.00
SPM5	**0.94**	**0.74**	**0.30**	**0.95**	0.00	**0.94**	0.16	**0.96**	0.00
SPM6	**0.81**	**0.73**	0.15	**0.84**	0.00	**0.83**	0.04	**0.84**	0.00
SPM7	**0.66**	**0.30**	**0.47**	0.00	**0.71**	**0.59**	**0.33**	**0.61**	**0.31**
SPM8	**0.70**	0.23	**0.57**	0.00	**0.75**	**0.60**	**0.42**	**0.61**	**0.41**
SPM9	**0.60**	0.20	**0.50**	0.00	**0.64**	**0.52**	**0.36**	**0.51**	**0.40**
SPM10	**0.79**	**0.43**	**0.47**	0.00	**0.84**	**0.73**	**0.31**	**0.75**	**0.30**
SPM11	**0.62**	-0.05	**0.75**	0.00	**0.66**	**0.44**	**0.58**	**0.44**	**0.64**
SPM12	**0.53**	**−0.36**	**0.99**	0.00	**0.57**	0.28	**0.80**	**0.31**	**0.72**
φ	-	**0.57**		**0.80**		0.00		0.00	

[1] Unidim = Unidimensional model. Unidim.M. = Unidimensional model with SPM4-SPM5 residual correlation estimated. BID.EFA = Bi-dimensional exploratory factor analysis. BID.CFA = Bi-dimensional confirmatory factor analysis. BEFA = Bi-factor exploratory factor analysis. BCFA = Bi-factor confirmatory factor analysis. All loadings over 0.30 are presented bolded. Phi = Inter-factor correlation. SPM4-SPM5 = Residual covariance between SPM4-SPM5 items. G = General Factor. S1= First specific factor. S2 = Second specific factor. Factor loadings with values > 0.30 appear bolded.

Table A5. Model fit indices for all tested models.

	Np	df	X^2	*p*	CFI	TLI	RMSEA	SRMR
Unidim.	24	54	221.75	0.00	0.94	0.93	0.08 (0.08–0.09)	0.11
BID.EFA/BEFA.	35	43	97.21	0.00	**0.98**	**0.97**	**0.05 (0.04–0.06)**	**0.06**
BID.CFA	25	53	163.39	0.00	0.96	0.96	0.07 (0.05–0.07)	0.09
BCFA	30	48	117.65	0.00	0.98	0.97	0.05 (0.04–0.07)	0.07

[1] Unidim = Unidimensional model. Unidim.M. = Unidimensional model with SPM4-SPM5 residual correlation estimated. BID.EFA = Bi-dimensional exploratory factor analysis. BID.CFA = Bi-dimensional confirmatory factor analysis. BEFA = Bi-factor exploratory factor analysis. BCFA = Bi-factor confirmatory factor analysis. Np = Estimated number of parameters. Df = degrees of freedom. ^{2} = Chi-square statistic. P = *p*-value associated with ^{2} test of fit. CFI = Comparative fit index. TLI= Tucker-Lewis index. RMSEA = Root Mean Square Error of Approximation (with 95% confidence interval in parenthesis). SRMS = Standardized Root Mean Square Residual. Best fit indices presented bolded and underlined. Model fit indices for the best fitting model appear bolded.

References

1. Raven, J.C. Standardization of Progressive Matrices. *Br. J. Med. Psychol.* **1941**, *19*, 137–150. [CrossRef]
2. Myszkowski, N.; Storme, M. A snapshot of g? Binary and polytomous item-response theory investigations of the last series of the Standard Progressive Matrices (SPM-LS). *Intelligence* **2018**, *68*, 109–116. [CrossRef]
3. Abad, F.J.; Colom, R.; Rebollo, I.; Escorial, S. Sex differential item functioning in the Raven's Advanced Progressive Matrices: Evidence for bias. *Pers. Indiv. Differ.* **2004**, *36*, 1459–1470. [CrossRef]
4. Lucio, P.S.; Cogo-Moreira, H.; Puglisi, M.; Polanczyk, G.V.; Little, T.D. Psychometric Investigation of the Raven's Colored Progressive Matrices Test in a Sample of Preschool Children. *Assessment* **2017**, 1–11. [CrossRef] [PubMed]
5. Walsch, N.A.; Nettelbeck, S.A.J.; Nicholas, R.B. Dimensionality of the Raven's Advanced Progressive Matrices: Sex Differences and Visuospatial Ability. *Pers. Individ. Differ.* **2016**, *100*, 157–166. [CrossRef]
6. Lohmann, D.F.; Lakin, J.M. Intelligence and Reasoning. In *The Cambridge Handbook of Intelligence*; Sternberg, R.J., Kaufman, S.B., Eds.; Cambridge University Press, Ltd.: New York, NY, USA, 2011; pp. 419–441.
7. Dillon, R.F.; Pohlmann, J.T.; Lohman, D. A Factor Analysis of Raven's Advanced Progressive Matrices Freed from Difficulty Factors. *Educ. Psychol. Meas.* **1981**, *41*, 1295–1302. [CrossRef]
8. Bors, D.A.; Stokes, T.L. Raven's Advanced Progressive Matrices: Norms for First-Year University Students and the Development of a Short Form. *Educ. Psychol. Meas.* **1998**, *58*, 392–398. [CrossRef]
9. Holzinger, K.; Swineford, F. The Bi-factor Method. *Psychometrika* **1937**, *2*, 41–54. [CrossRef]
10. Waller, N.G. Direct Schmid-Leiman Transformations and Rank-Deficient Loadings Matrices. *Psychometrika* **2017**, *83*, 858–887. [CrossRef] [PubMed]
11. Johnson, W.; Bouchard, T.J. The MISTRA data: Forty-two mental ability tests in three batteries. *Intelligence* **2011**, *39*, 82–88. [CrossRef]
12. Gignac, G.E. Raven's is not a pure measure of general intelligence: Implications for g factor theory and the brief measurement of g. *Intelligence* **2015**, *52*, 71–79. [CrossRef]
13. Muniz, M.; Gomez, C.; Pasian, S. Factor Structure of Raven's Coloured Progressive Matrices. *Psico-USF* **2016**, *21*, 259–272. [CrossRef]
14. Zeller, F.; Reiß, S.; Schweizer, K. Is the Item-Position Effect in Achievement Measures Induced by Increasing Item Difficulty. *Struct. Equ. Model.* **2019**, *24*, 745–754. [CrossRef]
15. Ren, X.; Wang, T.; Sun, S.; Deng, M.; Scheizer, K. Speeded testing in the assessment of intelligence gives rise to speed factor. *Intelligence* **2018**, *66*, 64–71. [CrossRef]
16. Sun, S.; Scheizer, K.; Ren, X. Item-Position Effect in Raven's Matrices: A Developmental Perspective. *J. Cogn. Dev.* **2019**, 1–10. [CrossRef]
17. Golino, H.F.; Shi, D.; Garrido, L.E.; Christensen, A.; Nieto, M.D.; Sadana, P.; Thiyagarajan, J.A. Investigating the performance of exploratory graph analysis and traditional techniques to identify the number of latent factors: A simulation and tutorial. *Psychol. Methods* **2019**, 1–20. [CrossRef]
18. Lubbe, D. Parallel Analysis with Categorical Variables: Impact of Category Probability Proportions on Dimensionality Assessment Accuracy. *Psychol. Methods* **2018**, *24*, 339–351. [CrossRef] [PubMed]

19. Marsh, H.; Morin, A.; Parker, P.; Kaur, G. Exploratory structural equation modeling: An integration of the best features of exploratory and confirmatory factor analysis. *Annu. Rev. Clin. Psychol.* **2014**, *10*, 85–110. [CrossRef] [PubMed]
20. Marsh, H.; Muthen, B.; Asparouhov, T.; Lüdke, O.; Robitzsch, A.; Morin, A.; Trautwein, U. Exploratory structural equation modelling, integrating CFA and EFA: Application to student's evaluations of university teaching. *Struct. Equ. Model.* **2009**, *16*, 439–476. [CrossRef]
21. Revelle, W.; Wilt, J. The general factor of personality: A general critique. *J. Res. Pers.* **2013**, *47*, 493–504. [CrossRef] [PubMed]
22. Timmerman, M.E.; Lorenzo-Seva, U. Dimensionality Assessment of Ordered Polytomous Items with Parallel Analysis. *Psychol. Methods* **2016**, *16*, 209–222. [CrossRef] [PubMed]
23. Garrido, L.E.; Abad, F.J.; Ponsoda, V. Are fit indices really fit to estimate the number of factors with categorical variables? Some cautionary findings via Monte Carlo simulation. *Psychol. Methods* **2016**, *21*, 93–111. [CrossRef] [PubMed]
24. Garrido, L.E.; Abad, F.J.; Ponsoda, V. A new look at Horn's parallel analysis with ordinal variables. *Psychol. Methods* **2013**, *18*, 454–474. [CrossRef] [PubMed]
25. Raiche, G.; Walls, T.; Magis, D.; Riopel, M.; Blais, J.G. Non-graphical Solutions for Cattell's Scree Test. *Methodology* **2013**, *9*, 23–29. [CrossRef]
26. Crawford, A.V.; Green, S.B.; Levy, R.; Lo, W.J.; Scott, L.; Svetina, D.; Thompson, M.S. Evaluation of parallel analysis methods for determining the number of factors. *Educ. Psychol. Meas.* **2010**, *70*, 885–901. [CrossRef]
27. Parry, D.H.; McArdle, J.J. An Applied Comparison of Methods for Least-Squares Factor Analysis of Dichotomous Variables. *Appl. Psychol. Meas.* **1991**, *15*, 35–46. [CrossRef]
28. Weng, L.J.; Cheng, C.P. Parallel Analysis with Unidimensional Binary Data. *Educ. Psychol. Meas.* **2005**, *65*, 697–716. [CrossRef]
29. Mulaik, S. *Foundations of Factor Analysis*, 2nd ed.; Chapman & Hall/CRC: Boca Raton, FL, USA, 2010.
30. Golino, H.F.; Epskamp, S. Exploratory Graph Analysis: A New Approach for Estimating the Number of Dimensions in Psychological Research. *PLoS ONE* **2017**, *12*, e0174035. [CrossRef] [PubMed]
31. Golino, H.F.; Demetriou, A. Estimating the Dimensionality of Intelligence like Data using Exploratory Graph Analysis. *Intelligence* **2017**, *62*, 54–57. [CrossRef]
32. Rodriguez, A.; Reise, S.P.; Haviland, M.G. Applying Bifactor Statistical Indices in the Evaluation of Psychological Measures. *J. Pers. Assess.* **2016**, *98*, 223–237. [CrossRef]
33. Rodriguez, A.; Reise, S.P.; Haviland, M.G. Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychol. Methods* **2016**, *21*, 137–150. [CrossRef] [PubMed]
34. Reise, S.P.; Scheines, R.; Widaman, K.; Haviland, M. Multidimensionality and Structural Coefficient Bias in Structural Equation Modelling: A Bifactor Perspective. *Educ. Psychol. Meas.* **2013**, *73*, 5–26. [CrossRef]
35. Reise, S.P.; Bonifay, W.; Haviland, M.G. Bifactor modelling and the evaluation of scale scores. In *The Wiley Handbook of Psychometric Testing: A Multidisciplinary Reference on Survey, Scale and Test Development*, 1st ed.; Irwing, P., Booth, T., Hughes, D.J., Eds.; John Wiley & Sons Ltd.: Hoboken, NJ, USA, 2018; Volume 2, pp. 675–707.
36. Garcia-Garzon, E.; Abad, F.J.; Garrido, L.E. Improving Bi-factor Exploratory Modelling: Empirical Target Rotation Based on Loading Differences. *Methodology* **2019**, *15*, 45–55. [CrossRef]
37. Mai, Y.; Zhang, Z.; Wen, Z. Comparing Exploratory Structural Equation Modeling and Existing Approaches for Multiple Regression with Latent Variables. *Struct. Equ. Model.* **2018**, *25*, 737–749. [CrossRef]
38. Abad, F.J.; Garcia-Garzon, E.; Garrido, L.E.; Barrada, J.R. Iteration of Partially Specified Target Matrices: Application to the Bi-Factor Case. *Multivar. Behav. Res.* **2017**, *52*, 416–429. [CrossRef] [PubMed]
39. Asparouhov, T.; Muthen, B. Exploratory Structural Equation Modeling. *Struct. Equ. Model.* **2009**, *16*, 397–438. [CrossRef]
40. Mansolf, M.; Reise, S.P. Exploratory Bifactor Analysis: The Schmid-Leiman Orthogonalization and Jennrich-Bentler Analytic Rotations. *Multivar. Behav. Res.* **2016**, *51*, 695–717. [CrossRef]
41. Lorenzo-Seva, U.; Ferrando, P.J. A General Approach for Fitting Pure Exploratory Bifactor Models. *Multivar. Behav. Res.* **2019**, *54*, 15–30. [CrossRef] [PubMed]
42. Lorenzo-Seva, U. Promin: A Method for Oblique Factor Rotation. *Multivar. Behav. Res.* **1999**, *34*, 347–365. [CrossRef]
43. Jennrich, R.I.; Bentler, P. Exploratory Bi-factor Analysis. *Psychometrika* **2011**, *76*, 537–549. [CrossRef]

44. Jennrich, R.I.; Bentler, P. Exploratory Bi-factor Analysis: The Oblique Case. *Psychometrika* **2012**, *77*, 442–454. [CrossRef] [PubMed]
45. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2019; Available online: https://www.R-project.org/ (accessed on 20 June 2019).
46. Allaire, J.J.; Xie, Y.; McPherson, J.; Luraschi, J.; Ushey, K.; Atkins, A.; Wickham, H.; Cheng, J.; Chang; Iannone, R. Rmarkdown: Dynamic Documents for R. R Package Version 1.12. 2019. Available online: https://rmarkdown.rstudio.com (accessed on 20 June 2019).
47. Aust, F.; Barth, M. Papaja: Prepare Reproducible APA Journal Articles with R Markdown. R Package Version 0.1.0.9842. 2018. Available online: https://github.com/crsh/papaja (accessed on 20 June 2019).
48. Epskamp, S.; Cramer, A.O.J.; Waldorp, L.J.; Schmittmann, V.D.; Borsboom, D. Qgraph: Network Visualizations of Relationships in Psychometric Data. *J. Stat. Softw.* **2012**, *48*, 1–19. Available online: http://jstatsoft.org&v48/i04/ (accessed on 20 June 2019).
49. Revelle, W. *Psych: Procedures for Personality and Psychological Research*; Northwestern University: Evanston, IL, USA, 2018; Available online: https://www.scholars.northwestern.edu/en/publications/psych-procedures-for-personality-and-psychological-research (accessed on 20 June 2019).
50. Golino, H. EGA: Exploratory Graph Analysis: Estimating the Number of Dimensions in Psychological Data. 2019. Available online: http://github.com/hfgolino/EGA (accessed on 20 June 2019).
51. Rosseel, Y. lavaan: An R Package for Structural Equation Modeling. *J. Stat. Softw.* **2012**, *48*, 1–36. Available online: http://www.jstatsoft.org/v48/i02 (accessed on 20 June 2019). [CrossRef]
52. Jorgensen, T.D.; Pornprasertmanit, S.; Schoemann, A.M.; Rosseel, Y. SemTools: Useful Tools for Structural Equation Modeling. R Package Version 0.5-1. 2018. Available online: https://CRAN.R-project.org/package=semTools (accessed on 20 June 2019).
53. Viladrich, C.; Angulo-Brunet, A.; Doval, E. A Journey Around Alpha and Omega to Estimate Internal Consistency Reliability. *Ann. Psychol.* **2017**, *33*, 755–782. [CrossRef]
54. Bernaards, C.A.; Jennrich, R.I. Gradient Projection Algorithms and Software for Arbitrary Rotation Criteria in Factor Analysis. *Educ. Psychol. Meas.* **2005**, *65*, 676–696. [CrossRef]
55. Little, T.D.; Rhemtulla, M.; Gibson, K.; Schoemann, A.M. Why the Items versus Parcels Controversy Needn't Be one. *Psychol. Methods* **2013**, *18*, 285–300. [CrossRef] [PubMed]
56. Estrada, E.; Román, F.J.; Abad, F.J.; Colom, R. Separating power and speed components of standardized intelligence measures. *Intelligence* **2017**, *61*, 159–168. [CrossRef]
57. Carroll, J.B. The effect of difficulty and chance success on correlations between items or between tests. *Psychometrika* **1945**, *10*, 1–19. [CrossRef]
58. DeMars, C.E. Revised Parallel Analysis with Nonnormal Ability and a Guessing Parameter. *Educ. Psychol. Meas.* **2019**, *79*, 151–169. [CrossRef] [PubMed]
59. Abad, F.J.; Sorrel, M.A.; Garcia, L.F.; Aluja, A. Modeling General, Specific, and Method Variance in Personality Measures: Results for ZKA-PQ and NEO-PI-R. *Assessment* **2018**, *25*, 959–977. [CrossRef]

Journal of
Intelligence

Article

Same Test, Better Scores: Boosting the Reliability of Short Online Intelligence Recruitment Tests with Nested Logit Item Response Theory Models

Martin Storme [1,2,*,†], Nils Myszkowski [3,†], Simon Baron [4] and David Bernard [4]

[1] IESEG School of Management, 59800 Lille, France
[2] LEM-CNRS 9221, 59800 Lille, France
[3] Department of Psychology, Pace University, New York, NY 10038, USA
[4] Assess First, 75000 Paris, France
[*] Correspondence: m.storme@ieseg.fr; Tel.: +33-320-54-20-44
[†] These authors contributed equally to this work.

Received: 29 April 2019; Accepted: 5 July 2019; Published: 10 July 2019

Abstract: Assessing job applicants' general mental ability online poses psychometric challenges due to the necessity of having brief but accurate tests. Recent research (Myszkowski & Storme, 2018) suggests that recovering distractor information through Nested Logit Models (NLM; Suh & Bolt, 2010) increases the reliability of ability estimates in reasoning matrix-type tests. In the present research, we extended this result to a different context (online intelligence testing for recruitment) and in a larger sample ($N = 2949$ job applicants). We found that the NLMs outperformed the Nominal Response Model (Bock, 1970) and provided significant reliability gains compared with their binary logistic counterparts. In line with previous research, the gain in reliability was especially obtained at low ability levels. Implications and practical recommendations are discussed.

Keywords: E-assessment; general mental ability; nested logit models; item-response theory; ability-based guessing

1. Introduction

With the development of the Internet, the assessment of job applicants is increasingly performed online, which facilitates large scale testing while reducing costs [1,2]. This recent trend has led to the creation of a new research field in psychometrics, referred to as *e-assessment* [2]. Considering the relevance of General Mental Ability (GMA) in predicting job performance [3], many e-assessment platforms have included tasks that aim at capturing it—such as logical series or logical reasoning matrices—in their online test batteries.

The assessment phase in e-recruiting poses very specific psychometric challenges. On the one hand, the assessment should ideally lead to a short-list of the best applicants [1,2]. The accuracy of the assessment is therefore a key issue in e-recruiting just like in it is in recruiting in general. On the other hand, the assessment phase cannot require from applicants that they take part in assessment processes that are too time consuming and too cognitively demanding. It is indeed not acceptable to extensively test people who have a relatively low chance of getting an interview. Perceived unfairness of the recruitment process has been shown to have a negative impact on the image of the recruiting company, which can lead to negative word of mouth and/or intentions not to complete the recruitment process [4,5]. The challenge that is inherent to e-assessment in a recruitment context is essentially the challenge of short psychometric measures, which is to extract as much information as possible from short instruments.

Extracting reliable information from short tests remains a real challenge from a measurement perspective [6]. Hopefully, psychometricians have allies in this challenging endeavour, such as Item Response Theory (IRT) modeling, which often allows them to extract more information from short psychometric tools than Classical Test Theory (CTT). Originally suggested for multiple-choice items by Bock [7], one way that researchers can take advantage of the IRT framework in logical series or matrices tests consists in extracting information from which incorrect responses were selected. This approach is based on the premise that when a test taker selects a wrong response option out of a set of wrong response options, the choice of the wrong response option can carry information about the ability of the test taker. Further, recent developments [8] applied to progressive matrices have suggested recovering additional information from distractor responses through Nested Logit Models (NLM) [9], and have indicated that such models may be more appropriate than Bock's [7] Nominal Response Model in logical reasoning tests, but also than traditional binary IRT models [8]. In this research, recovering information from the choice of distractors has provided significant gains in reliability in comparison with not recovering such information and using traditional binary logistic models.

Currently, no study has investigated whether applying this approach in the field of recruitment would lead to gains in reliability. Yet, taking an online GMA test as part of a recruitment process is in several ways different from taking a GMA test for an experiment in the lab. There is reasonable evidence to suspect that such differences could affect the way distractors are processed by test takers, which could possibly jeopardize the very idea of recovering psychometric information from distractors. In the present article, our main aim is to extend and conceptually replicate previous research on students and in laboratory conditions [8] to online personnel pre-selection contexts, by testing whether the modeling strategies previously suggested are able, even in this context, to provide tangible gains in reliability. The effort of conducting conceptual replications in the field is crucial in psychology to rule out the possibility that a laboratory finding is too weak to be relevant in contexts that are less tightly controlled [10].

1.1. Binary Item Response Theory Models

Item Response Theory (IRT) has traditionally helped psychometricians improve the reliability of the ability estimates obtained with short intelligence measures [8,11]. IRT provides a framework that has indeed been shown to improve the reliability of measurement compared to the Classical Test Theory (CTT) approach [12]. While CTT assumes that all items are linked to the latent trait in a similar fashion, IRT assumes that each item is linked to the the latent trait in a unique manner [13]. The aim of IRT is to model the probability of a response to an item as a function of the latent trait or ability of the test taker, traditionally with a non-linear function of the latent trait that is unique for each item. In the case of binary responses, the non-linear function is, frequently, the logistic function. Because of the flexibility of its parametrization in comparison with CTT, IRT allows for the accounting of a variety of testing phenomena and extracting information that is relevant in the context of GMA assessment [8].

GMA tests, such as progressive matrices or logical series, usually contain one correct answer option and several incorrect answer options—which are often referred to as distractors. Although the response dataset is thus polytomous, it is typical to recode the dataset by collapsing the distractor responses together, which yields a dichotomous success/failure variable format. The binary IRT approach generally consists in modeling these dichotomous responses using a logistic function of the latent ability and a set of item parameters representing various item characteristics (difficulty, discrimination, etc.).

The simplest IRT models, including only one parameter and referred to as One-Parameter Logistic (1PL) models, characterize items by their level of difficulty only. The difficulty parameter corresponds to the level of the latent trait for which the slope of the function linking the ability and the probability to select the correct response option reaches its maximum—in other words, the ability level where the discrimination of the item is at its maximum. The model is often extended with another parameter—discrimination—leading to Two-Parameter Logistic (2PL) models. Such models not only take into account the difficulty of an item, but also its ability to discriminate between ability levels. The discrimination parameter corresponds to the strength of relationship between the ability and the probability to select the correct response option. Three-Parameter Logistic (3PL) models add to previous models a variable lower asymptote in the relation between the ability and the probability to select the correct response option. In the context of IRT, the lower asymptote corresponds to the probability to select the correct answer to a given item by guessing it. Therefore, 3PL models allow to characterize items regarding the extent to which they are susceptible to correct guessing. A fourth parameter is included in 4-Parameter Logistic (4PL) models, which corresponds to a variable upper asymptote in the relation between the ability and the probability to select the correct response option. In the context of IRT, the upper asymptote corresponds to the probability of responding incorrectly to an item in spite of having a level of ability that should normally lead to providing the right answer. This parameter allows the modelling of the phenomenon of inattention or slipping. Although 4PL models are used less frequently than 1, 2, and 3PL models, they have been shown to correct adequately for careless mistakes and to improve measurement efficiency [14,15].

Although binary IRT is able to model phenomena that appear in matrix-type reasoning tasks, even models that include guessing fail to account for the possibility that choosing a distractor over another one could be related to the respondent's ability—a phenomenon often described as as ability-based guessing [16]. Indeed, the lower asymptote parameter of the 3PL and 4PL models account for the probability of correctly guessing, but what distractor is chosen when an examinee uses a guessing strategy is not considered—all distractor responses are still collapsed together as incorrect. Yet, if one considers that the guessing process is related to the ability, then the outcome of this process—the distractor chosen—can contain information about the ability that binary models fail to recover.

1.2. Recovering Distractor Information

In matrix-type or logical series type tests, distractors are usually designed in a way that they are only partially in line with the set of rules that structures the logical series. For example, if three rules are structuring the progression of a logical series, the correct response option will respect all three of them, but frequently a distractor could respect only two, while another may respect one or even none of them. In this example, a distractor that respects two out of three rules could be considered as a better response option than a distractor that would only respect one out of three rules, although both are ultimately incorrect response options. As a consequence, the wrong response options that are selected by test takers are usually not equivalent in (in)correctness, and thus could carry information about their ability [17].

1.2.1. The Nominal Response Model

A traditional approach to recovering information from distractors is to fit the nominal data with Bock's [7] Nominal Response Model (NRM). This model is essentially a multinomial adaptation of the 2PL model, where the probability P_{iv} that an examinee j chooses a category v—which could be the correct response or a distractor—among the m_i possible responses for item i is modeled as a function of the examinee's ability θ_j, an intercept item-category parameter ζ_{iv} and a slope item-category parameter λ_{iv}, as well as the item-category parameters of all other categories, such as:

$$P_{iv}\left(\theta_j\right) = \frac{e^{\zeta_{iv}+\lambda_{iv}\theta_j}}{\sum_{k=1}^{m_i} e^{\zeta_{ik}+\lambda_{ik}\theta_j}} \tag{1}$$

A way to interpret this model is to essentially consider each category as having a propensity $e^{\zeta_{iv}+\lambda_{iv}\theta_j}$ and the probability of selecting a category depends on the category's propensity over the total of all category propensities. When applied to multiple choice items, a consequence of this is that the Nominal Response Model's formulation is mathematically consistent with a response process where all response categories compete with one another and where, depending on the examinee's ability, one category would dominate in propensity the others, and result in the examinee responding (more probably) in favor of that category [9]. But, as we later discuss, this representation of the response process may not be in line with all multiple-choice tests, especially in the case of logical reasoning matrices or logical series.

1.2.2. Nested Logit Models

In certain multiple-choice tests, in order to respond, the examinee is supposed to consider a stimulus (for example, in matrix-type tests, the incomplete matrix), from which a rule should be extracted and used to find the completing element. In such cases, it can be questioned whether examinees put into competition the different response options right away—a process that would ideally be modeled by the NRM. Instead, it could be that they first focus on understanding the stimulus (the matrix, or the beginning of the series) to find the correct response (regardless of what the response options are). From that process, two situations may arise—either they have understood the rule correctly and found the correct response—in that case, the distractors are not really considered as viable options and the correct response is selected—or they have not—and in that case the response options are put in competition in the guessing strategy.

Such a sequential process was described by Suh and Bolt [9] as a motivation to develop a new class of item-response models for multiple-choice questions where this double process could be considered: Nested Logit Models (NLM). NLMs have been designed to model situations in which the response choice possesses a nested structure, that is when the final choice of a response option is made through a sequential process.

NLMs attempt to approximate the response probabilities that occur from this sequential process and the two models that best describe each step into a single model. NLMs have two levels that separate the response options in two nests. At a higher level (level 1), the model distinguishes the choice of the correct response option versus the choice of any incorrect response option, which can be achieved with a binary logistic IRT model (e.g. the 2PL, 3PL or 4PL model). At a lower level (level 2), the model distinguishes the probability of selecting one particular distractor (as opposed to another one) as the product of the probability of selecting any distractor (which is the complement of the probability earlier modeled with the level 1 part) and a probability modeled using the propensities of each distractor—which is similar to a Nominal Response Model of the distractors.

To summarize, using the 4-Parameter NLM (4PNL) as an example for at level 1, the probability $P(x_{ij} = u|\theta_j)$ that the jth person selects the correct response (category u) to the ith item, depends on their ability θ_j and item parameters α_i (discrimination/slope), β_i (difficulty/intercept), γ_i (lower asymptote) and δ_i (upper asymptote), such as:

$$P(x_{ij} = u|\theta_j) = \gamma_i + \frac{\delta_i - \gamma_i}{1 + e^{-\left(\beta_i + \alpha_i\theta_j\right)}} \tag{2}$$

Similar to binary logistic models, the 3-Parameter Nested Logit (3PNL) model is a constrained 4PNL where δ_i is fixed, generally (and throughout in this paper) to 1, such as:

$$P(x_{ij} = u|\theta_j) = \gamma_i + \frac{1 - \gamma_i}{1 + e^{-\left(\beta_i + \alpha_i\theta_j\right)}} \tag{3}$$

Further, the 2-Parameter Nested Logit (2PNL) is a constrained 3PNL where γ_i is fixed, generally (and throughout in this paper) to 0, such as:

$$P(x_{ij} = u|\theta_j) = \frac{1}{1 + e^{-\left(\beta_i + \alpha_i\theta_j\right)}} \tag{4}$$

At level 2, which models the distractor responses, the probability $P(x_{ij} = v|\theta_j)$ that the examinee selects the distractor category v among the m_i possible distractor responses is modeled as the product of the probability of responding incorrectly $1 - P(x_{ij} = u|\theta_j)$ and the probability that the examinee selects the distractor conditional upon an incorrect response. The latter is in fact similar to a Nominal Response model, where distractor responses have propensities that are a function of the ability θ_j, intercept ζ_{iv} and slope λ_{iv}. The resulting distractor model for the probability $P\left(U_{ij} = 0, D_{ijv}|\theta_j\right)$ that person j selects distractor v for item i is thus given by:

$$P(x_{ij} = v|\theta_j) = \left[1 - P(x_{ij} = u|\theta_j)\right]\left[\frac{e^{\zeta_{iv}+\lambda_{iv}\theta_j}}{\sum_{k=1}^{m_i} e^{\xi_{ik}+\lambda_{ik}\theta_j}}\right] \tag{5}$$

Using the level 1 4PL model in Equation (2), the distractors-model results in the 4PNL model to:

$$P(x_{ij} = v|\theta_j) = \left[1 - \left(\gamma_i + \frac{\delta_i - \gamma_i}{1 + e^{-\left(\beta_i + \alpha_i\theta_j\right)}}\right)\right]\left[\frac{e^{\zeta_{iv}+\lambda_{iv}\theta_j}}{\sum_{k=1}^{m_i} e^{\xi_{ik}+\lambda_{ik}\theta_j}}\right] \tag{6}$$

Using the level 1 3PL model in Equation (3), the distractors-model results in the 3PNL model to:

$$P(x_{ij} = v|\theta_j) = \left[1 - \left(\gamma_i + \frac{1 - \gamma_i}{1 + e^{-\left(\beta_i + \alpha_i\theta_j\right)}}\right)\right]\left[\frac{e^{\zeta_{iv}+\lambda_{iv}\theta_j}}{\sum_{k=1}^{m_i} e^{\xi_{ik}+\lambda_{ik}\theta_j}}\right] \tag{7}$$

Using the level 1 2PL model in Equation (4), the distractors-model results in the 2PNL model to:

$$P(x_{ij} = v|\theta_j) = \left[1 - \frac{1}{1 + e^{-\left(\beta_i + \alpha_i\theta_j\right)}}\right]\left[\frac{e^{\zeta_{iv}+\lambda_{iv}\theta_j}}{\sum_{k=1}^{m_i} e^{\xi_{ik}+\lambda_{ik}\theta_j}}\right] \tag{8}$$

An important distinction to note between the models of this class and the Nominal Response Model is that, in the NLM, the probability of a correct response is not directly affected by the propensities towards the different distractors, but the probability to select the distractors is conditional upon the probability of a correct (or rather, incorrect) response. In contrast, in the Nominal Response Model, the propensities towards all response categories—correct response and distractors alike—all affect one another.

To illustrate NLM, we present in Figure 1 the item-category characteristic curves for an item of the test studied in this very paper.

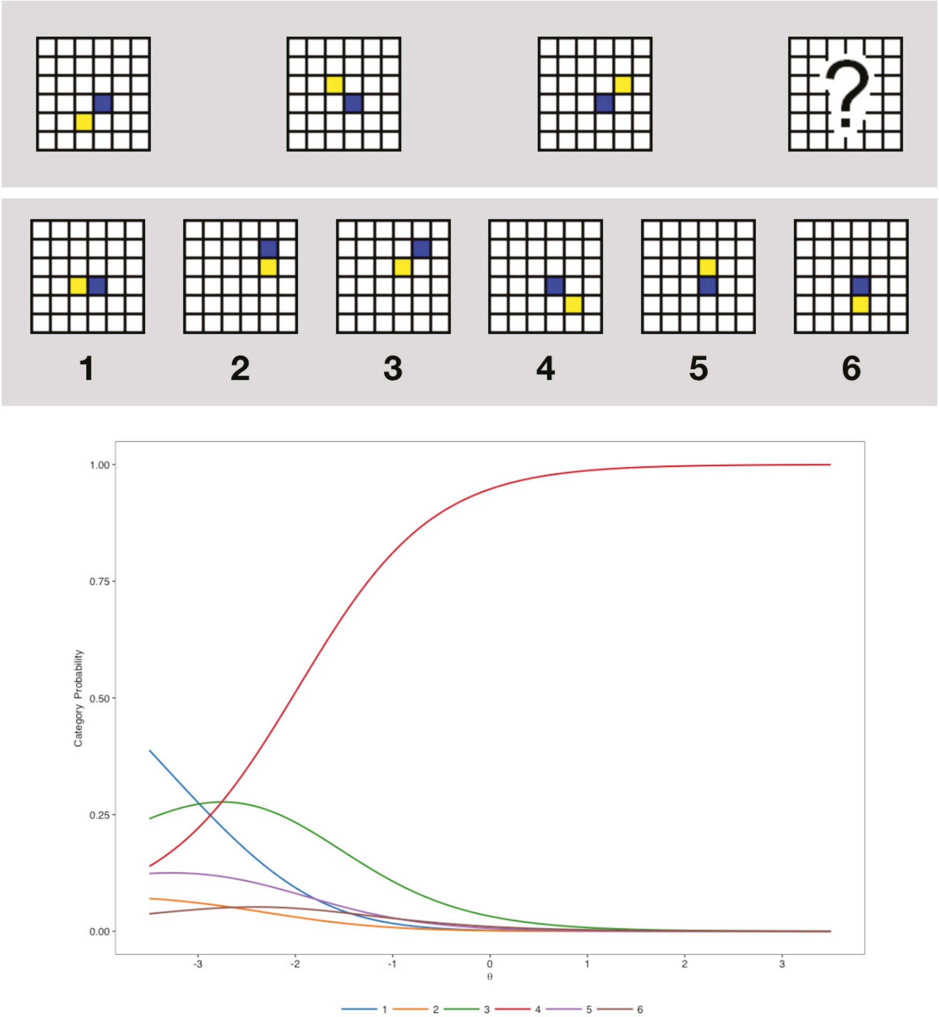


Figure 1. An example item of the GF20 (**top**) and the associated category characteristic curves as estimated by the 3-Parameter Nested Logit model (**bottom**). The correct response (4) is increasingly probable as θ_j increases. However, the response category 3—which is the only distractor response where the blue and the yellow squares are (correctly) not adjacent—would be more probably selected by individuals with low abilities ($\theta_j \approx -2.7$), while the category 1 would be more probably selected by individuals with even lower abilities ($\theta_j < -3$)—thus showing that the choice of distractor may be informative of θ_j.

1.3. The Aim of This Study

Although originally, Nominal Response Models were considered as a way to recover information from multiple-choice tests, recent research suggests that, in the case of matrix or series-type GMA tests, NLM may better fit the norminal-level data than the NRM and provide significant reliability gains in comparison with binary logistic models. In particular, Myszkowski and Storme [8] have shown that, on the last series of the Standard Progressive Matrices [18], (1) using NLM provided a better fit than Nominal Response Models to the nominal level data, and (2) NLMs allowed significant reliability gains when estimating ability.

Yet, however promising, this result has only been observed with one GMA test and has only been used on a convenience sample of undergraduates, in a low-stakes situation. This study aims at bridging this gap by replicating this result on another short GMA test, with higher stakes, and in a context that would be particularly interested in these reliability gains: Online recruitment.

The conditions under which job applicants take GMA tests are indeed very different from the conditions in which research participants take similar tests in the lab as part of a typical research study. For example, in a recruitment context, the stakes are higher in comparison with taking the test in a lab. Previous research on the effect of pressure on cognitive processes when taking intelligence tests has shown that when under pressure, working memory is busy processing intrusive thoughts which can have in turn a negative impact on performance [19,20]. It is possible that this phenomenon also affects the way distractors are processed and lead to different processing of response options. When under pressure, test takers who fail at identifying the rule that structures the progression of the series might experience high levels of stress and fail at comparing efficiently distractors to identify the best of the incorrect response options. As a consequence, in the context of the online assessment of job applicants, the choice of distractors might carry little information about the ability of test takers. If this is the case, NLMs should not lead in this context to gains in empirical reliability compared with binary models.

Furthermore, the fact that job applicants usually take online tests in their own time leads to less standardized testing contexts. Compared with the relatively controlled and quiet conditions of a lab, there might be more attentional perturbations in the environment, which might induce a shallower processing of the wrong response options. Consequently, it is possible that in the context of e-assessment, the choice of distractors is not so much reflective of the ability of the test taker, which could hinder the potential gains from NLMs.

The aim of the present study is to test whether the findings of Myszkowski and Storme [8]—obtained in a low psychological pressure and controlled context—can be replicated and generalized to an assessment situation characterized by more psychological pressure and less standardization, as well as a different test.

2. Method

2.1. Participants and Procedure

The sample consisted of 2949 French job applicants (2084 Men, 865 Women, M_{age} = 36.88, SD_{age} = 8.66) who responded to a logical series test that aims at measuring GMA online. The examinees responded using an e-assessment application presented in their web browser. As it is common in e-assessment, it can be expected that the standardization regarding when and where the test was taken was relatively low as job applicants were free to take the test at the time and at the place that was the most convenient for them. Of the participants, 40.96% had a master (or higher) degree, 23.64% of participants had a bachelor degree, the remaining applicants had less than a bachelor degree.

2.2. Instrument

The test under investigation—the GF20—comprises 20 incomplete logical series presented each with six response options to complete the missing part, including one correct answer that can be deducted from the application of logical rules. Each logical series consists of a 4 by 1 matrix with colored cells moving progressively on a grid according to simple geometric rules—such as translations and rotations. The 20 items that are comprised in the final test were designed and pre-tested to discriminate different levels of ability. An item example is provided in Figure 1. Except for instructions participants to complete the series, the test only included non-verbal and non-numerical content. No time limit was provided to applicants to take the test. It took them on average 21.30 min to complete the 20 items (SD = 9.78). Items were presented one by one. Participants were instructed to provide an response to each item before they could move on to the next item, and were not able to go back.

The CTT-based reliability estimates—computed using the R package "semtools" [21] from a unidimensional model fit with the package "lavaan" [22]—of the GF20 were satisfactory, as Cronbach's α was 0.831, Raykov's ω congeneric reliability was 0.834 and McDonald's ω_h reliability was 0.822.

2.3. Binary IRT Modeling

2.3.1. Model Estimation

All binary IRT models—the 1-Parameter Logistic (1PL), 2-Parameter Logistic (2PL), 3-Parameter Logistic with free lower asymptote (3PL), and 4-parameter logistic (4PL) models—were estimated using an Expectation-Maximization (EM) algorithm with the R package "mirt" [23]. All models successfully converged. Nevertheless, the information matrix of the 4PL model could not be inverted in order to compute the parameter standard errors—decreasing the convergence tolerance and changing the estimation method did not solve this issue—which may be a sign that the estimates were unstable. Item characteristic curve plots, which, for binary models, present the expected probability of a correct response as a function of the latent ability θ_j were plotted using the package for R "jrt" [24]. To keep the paper concise, only models with appropriate fit were plotted.

2.3.2. Model Fit

The fit of the models were then compared on Likelihood Ratio Tests (LRT) the model's corrected Akaike Information Criterion (AICc). For the former, p values below 0.05 were used to indicate a significantly improved fit from using the more complex (least constrained) model as opposed to the least complex (most constrained) model. For the latter, a smaller AICc indicates a better (more parsimonious) model fit.

In addition, absolute model fit indices were obtained by using limited information Goodness-of-Fit statistics [25] as implemented in "mirt." As usual—although more frequently seen in Structural Equation Modeling—and similar to the original study of Myszkowski and Storme [8], we used as absolute model fit indices the Comparative Fit Index (CFI) and Tucker-Lewis Index (TLI), with thresholds of 0.95, along with the Standardized Root Mean Square Residual (SRMR) with a threshold of 0.08, and the Root Mean Square Error of Approximation (RMSEA) with a threshold of 0.06.

2.3.3. Reliability

Since the aim of this paper is to extend and replicate the finding that NLM provides an increase in measurement accuracy in logical GMA tests—as found with the Raven's progressive matrices [8]—quantifying measurement accuracy is key. Measurement accuracy is represented by several statistics in IRT, especially information, standard error of measurement and reliability, which are mathematical transformations of one another. Because reliability is a familiar metric for most researchers—in both CTT and IRT—is conveniently bounded by 0 and 1, and is the metric chosen in the article that this study attempts to replicate, it was chosen in this study. However, it should be noted that the conclusions reached about reliability here are also extendable to information and standard errors.

Similar to the original study, reliability functions were plotted for the 2PL, 3PL and 4PL models, overlayed with their Nested Logit counterparts. In addition, and also similar to the original study, marginal estimates of empirical reliability were computed [26]. The estimate of empirical reliability reported corresponds to the reliability of the θ_j scores, averaged across all cases j.

2.4. Nominal and Nested Logit IRT Models

2.4.1. Model Estimation

The models for nominal data—the Nominal Response (NR), the 2-Parameter Nested Logit (2PNL), the 3-Parameter Nested Logit (3PNL) an the 4-Parameter Nested Logit (4PNL) models—were estimated using the package "mirt" [23] using an EM algorithm. All models converged successfully. However, as with the binary models, the information matrix of the 4PNL model could not be inverted in order to compute the parameter standard errors, which may be a sign that the estimates were unstable. As for the binary models, item category curve plots, which present the expected probability of selecting a category as a function of the latent ability θ_j were computed using "jrt" [24]. Again, to keep the paper concise, only models with appropriate fit were reported.

2.4.2. Model Fit

Similar to the binary models, Likelihood Ratio Tests were used to compare the different nominal models. However, only the 2PNL, 3PNL and 4PNL models are nested with one another, and thus only they allow the use of Likelihood Ratio Tests when comparing them. The AICcs of all models were computed, and the AICc was used to compare the Nominal Response model with the other models.

Polytomous models are largely more heavily parametrized than binary models, which, in some cases, prevents to compute limited information Goodness-of-Fit statistics, such as in Myszkowski and Storme [8]—thereby limiting model fit estimations. However, in this case, because of the larger sample size than in Myszkowski and Storme [8], we were able to compute them, and used the same indices and thresholds earlier discussed for the binary models.

2.4.3. Reliability

Similar to the binary models, we also computed the reliability functions of the NLMs, which were plotted as an overlay of the reliability functions of their respective binary counterparts (i.e., 2PL and 2PNL, 3PL and 3PLN, 4PL and 4PLN)—thereby facilitating visual comparisons. We also computed the empirical reliability of each model averaged across cases as an estimate of marginal reliability.

As one of the aims of this study is to examine potential gains in reliability from using NLMs as opposed to their binary counterparts, we computed the reliability gain $\Delta r_{xx'}$ between models by computing their difference. Similar to the original study and other previous studies [8,15], we used bootstrapping to obtain a Wald's z test (based on the bootstrapped standard error) and 95% Confidence Intervals for the reliability gains.

3. Results

3.1. Binary IRT Models

The model fit indices of all binary models are reported in Table 1. The 2PL, 3PL and 4PL models all had satisfactory fit, with the 4PL model providing the best fit. The 4PL model fitted significantly better than the 3PL model ($\Delta\chi^2 = 167.405$, $\Delta df = 20$, $p < 0.001$), which fitted significantly better than the 2PL model ($\Delta\chi^2 = 519.018$, $\Delta df = 20$, $p < 0.001$), which fitted significantly better than the 1PL model ($\Delta\chi^2 = 1100.652$, $\Delta df = 19$, $p < 0.001$).

Table 1. Model fit of the binary models.

Model	χ^2	df	p	CFI	TLI	RMSEA	AICc
1-Parameter Logistic	2462.597	189	<0.001	0.913	0.913	0.064	58,244.74
2-Parameter Logistic	1069.812	170	<0.001	0.966	0.962	0.042	57,182.90
3-Parameter Logistic	251.3807	150	<0.001	0.996	0.995	0.015	56,705.29
4-Parameter Logistic	196.2342	130	<0.001	0.997	0.996	0.013	56,579.87

As they were the best two fitting models and provided very similar absolute fit indices, we present the item characteristic curves of both the 2PL, 3PL and 4PL models respectively in Figures 2–4. Their similarity and the relatively high low asymptotes for the 4PL model—for the 3PL, they are fixed to 1—are in line with the fact that the two models provided similar fit.

The parameter estimates (along with standard errors for the 2PL and 3PL) of the 2PL, 3PL, and 4PL models are presented respectively in Table 2.

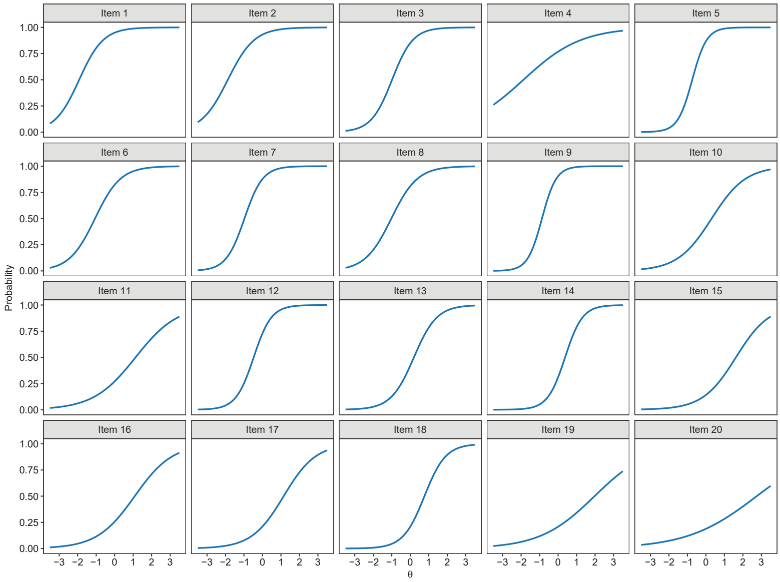


Figure 2. Item characteristic curve plots of the 2-Parameter Logistic Model.

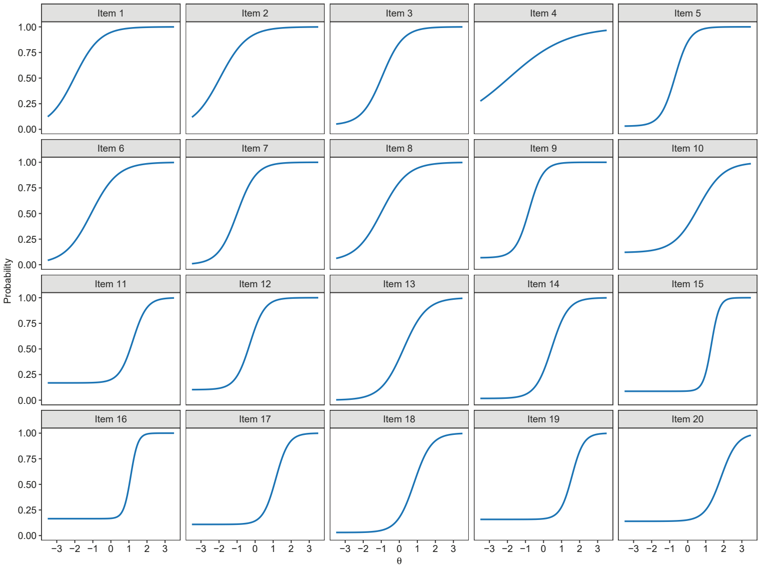


Figure 3. Item characteristic curve plots of the 3-Parameter Logistic Model.

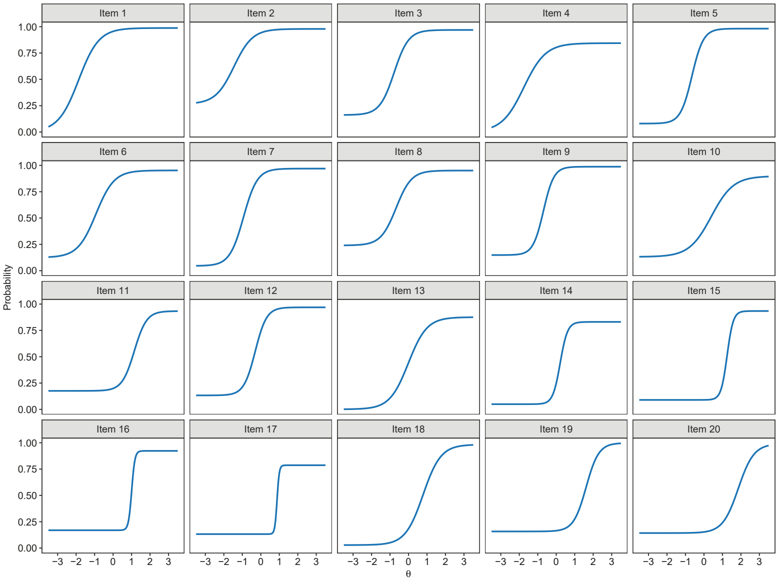


Figure 4. Item characteristic curve plots of the 4-Parameter Logistic Model.

The marginal estimates of empirical reliability for all the binary models were satisfactory and close to the CTT-based estimates earlier reported, as they were 0.833 for the 1PL model, 0.849 for the 2PL model, 0.868 for the 3PL model and 0.873 for the 4PL model.

Table 2. Item parameters of binary logistic models.

Item	**1PL Model**	**2PL Model**		**3PL Model**			**4PL Model**			
	β_i	α_i	β_i	α_i	β_i	logit(γ_i)	α_i	β_i	logit(γ_i)	logit(δ_i)
Item 1										
Estimate	2.783	1.527	2.930	1.417	2.855	−4.089	1.821	3.391	0.002	0.988
Standard error	0.072	0.105	0.113	0.104	0.157	6.671				
Item 2										
Estimate	2.569	1.391	2.605	1.326	2.575	−5.002	1.999	2.898	0.266	0.979
Standard error	0.068	0.094	0.096	0.087	0.104	6.298				
Item 3										
Estimate	1.513	1.767	1.740	1.735	1.633	−3.170	2.558	1.988	0.161	0.969
Standard error	0.055	0.093	0.078	0.155	0.136	2.191				
Item 4										
Estimate	1.454	0.640	1.198	0.619	1.185	-5.598	1.697	2.979	0.002	0.844
Standard error	0.055	0.054	0.048	0.051	0.057	6.083				
Item 5										
Estimate	1.291	2.543	1.878	2.462	1.719	−3.465	3.222	2.097	0.080	0.982
Standard error	0.053	0.132	0.099	0.179	0.102	1.223				
Item 6										
Estimate	1.475	1.442	1.535	1.362	1.477	−4.939	2.028	1.880	0.125	0.952
Standard error	0.055	0.078	0.067	0.083	0.089	6.479				
Item 7										
Estimate	1.588	2.015	1.966	1.891	1.884	−6.457	2.667	2.501	0.045	0.969
Standard error	0.056	0.106	0.089	0.097	0.084	6.179				
Item 8										
Estimate	1.404	1.412	1.448	1.389	1.365	−3.355	2.415	1.622	0.240	0.951
Standard error	0.054	0.077	0.064	0.141	0.178	3.531				
Item 9										
Estimate	1.542	2.575	2.245	2.593	2.061	−2.619	3.540	2.450	0.148	0.987
Standard error	0.055	0.138	0.111	0.216	0.118	0.751				
Item 10										
Estimate	−0.372	1.085	−0.335	1.438	−0.852	−2.002	1.683	−0.669	0.131	0.898
Standard error	0.050	0.059	0.046	0.149	0.172	0.285				

Table 2. *Cont.*

Item	1PL Model	2PL Model		3PL Model			4PL Model			
	β_i	α_i	β_i	α_i	β_i	logit(γ_i)	α_i	β_i	logit(γ_i)	logit(δ_i)
Item 11										
Estimate	−1.137	0.878	−0.991	2.603	−3.188	−1.597	3.013	-3.462	0.176	0.934
Standard error	0.052	0.055	0.048	0.327	0.400	0.093				
Item 12										
Estimate	0.762	2.078	0.991	2.440	0.697	−2.171	3.235	0.976	0.133	0.969
Standard error	0.051	0.101	0.070	0.177	0.099	0.276				
Item 13										
Estimate	−0.313	1.612	−0.316	1.577	−0.368	−7.978	1.988	0.021	0.001	0.876
Standard error	0.049	0.079	0.054	0.076	0.055	6.064				
Item 14										
Estimate	−0.662	2.121	−0.802	2.191	−0.992	−4.072	5.037	−1.078	0.049	0.831
Standard error	0.050	0.105	0.067	0.146	0.106	0.616				
Item 15										
Estimate	−1.926	1.113	−1.807	4.520	−5.921	−2.352	6.060	−7.593	0.090	0.934
Standard error	0.059	0.068	0.066	0.608	0.762	0.090				
Item 16										
Estimate	−1.186	0.981	−1.064	5.056	−5.569	−1.622	11.675	−11.750	0.169	0.923
Standard error	0.053	0.058	0.051	0.754	0.841	0.071				
Item 17										
Estimate	−1.399	1.153	−1.321	2.815	−3.228	−2.099	16.917	−14.883	0.132	0.787
Standard error	0.054	0.065	0.057	0.311	0.353	0.111				
Item 18										
Estimate	−1.192	1.736	−1.330	2.104	−1.729	−3.465	2.079	−1.636	0.028	0.983
Standard error	0.053	0.089	0.068	0.156	0.143	0.343				
Item 19										
Estimate	−1.603	0.678	−1.335	3.085	−4.858	−1.673	2.931	−4.739	0.158	0.998
Standard error	0.056	0.054	0.050	0.520	0.759	0.077				
Item 20										
Estimate	−1.808	0.532	−1.463	2.248	−4.156	−1.817	2.369	−4.404	0.143	0.990
Standard error	0.058	0.053	0.050	0.372	0.580	0.089				

3.2. Nominal Models

The model fit indices of all nominal models are reported in Table 3. Although the Nominal Response model provided a borderline acceptable fit, it was, as hypothesized, outperformed by all the NLMs, which all presented satisfactory fit. The 4PNL model fitted significantly better than the 3PNL model ($\Delta\chi^2 = 82.624$, $\Delta df = 20$, $p < 0.001$), which fitted significantly better than the 2PNL model ($\Delta\chi^2 = 541.102$, $\Delta df = 20$, $p < 0.001$).

The item category curve plots of the 2PNL, 3PNL and the 4PNL are respectively presented in Figures 5–7. Their model estimates as well as standard errors are presented respectively in Tables 4–6.

Table 3. Model fit of the nominal and nested logit models.

Model	χ^2	df	p	CFI	TLI	RMSEA	AICc
Nominal Response	178.0345	90	<0.001	0.972	0.941	0.018	134,347.1
2-Parameter Nested Logit	177.3853	90	<0.001	0.978	0.958	0.018	133,725.8
3-Parameter Nested Logit	126.1003	70	<0.001	0.986	0.965	0.016	133,231.1
4-Parameter Nested Logit	104.8853	50	<0.001	0.986	0.952	0.019	133,195.5

Table 4. Item parameters of the 2PNL model.

Item	Correct Response		Distractors							
	α_i	β_i	$\lambda_{i,1}$	$\lambda_{i,2}$	$\lambda_{i,3}$	$\lambda_{i,4}$	$\delta_{i,1}$	$\delta_{i,2}$	$\delta_{i,3}$	$\delta_{i,4}$
Item 1										
Estimate	1.549	2.954	0.426	1.067	0.744	1.327	−0.312	2.878	1.248	1.771
Standard error	0.103	0.113	0.535	0.341	0.383	0.382	0.805	0.523	0.579	0.554
Item 2										
Estimate	1.397	2.614	−1.189	−0.297	−0.123	−0.442	−3.219	−1.154	−1.509	−1.669
Standard error	0.091	0.096	0.357	0.194	0.229	0.231	0.548	0.235	0.266	0.292

Table 4. *Cont.*

Item	Correct Response		Distractors							
	α_i	β_i	$\lambda_{i,1}$	$\lambda_{i,2}$	$\lambda_{i,3}$	$\lambda_{i,4}$	$\delta_{i,1}$	$\delta_{i,2}$	$\delta_{i,3}$	$\delta_{i,4}$
Item 3										
Estimate	1.747	1.736	−0.915	−0.392	−0.914	−0.559	−1.147	−1.094	−1.369	−0.593
Standard error	0.089	0.077	0.183	0.193	0.195	0.163	0.205	0.197	0.222	0.167
Item 4										
Estimate	0.646	1.200	0.828	2.623	2.413	0.132	2.746	6.883	4.011	0.168
Standard error	0.053	0.048	0.650	0.646	0.671	0.821	1.150	1.122	1.136	1.476
Item 5										
Estimate	2.417	1.819	0.648	0.180	0.386	−0.077	1.893	0.351	1.343	−0.259
Standard error	0.119	0.093	0.215	0.255	0.221	0.279	0.260	0.313	0.270	0.355
Item 6										
Estimate	1.412	1.524	−1.426	−1.127	−1.334	−1.266	−2.516	−2.816	−2.308	−2.755
Standard error	0.075	0.066	0.205	0.244	0.194	0.231	0.245	0.283	0.226	0.274
Item 7										
Estimate	1.945	1.933	−0.373	−0.506	−1.447	−1.262	−0.445	−0.617	−1.647	−1.845
Standard error	0.098	0.085	0.171	0.179	0.218	0.237	0.171	0.184	0.267	0.291
Item 8										
Estimate	1.425	1.457	−0.868	−0.485	−0.010	−1.611	0.098	−0.573	0.507	−2.450
Standard error	0.075	0.064	0.163	0.194	0.153	0.288	0.161	0.189	0.136	0.384
Item 9										
Estimate	2.435	2.170	0.354	0.517	0.485	0.233	1.225	0.780	0.208	1.577
Standard error	0.123	0.103	0.244	0.270	0.307	0.230	0.303	0.325	0.365	0.291
Item 10										
Estimate	1.090	−0.336	0.327	0.472	0.248	1.177	0.701	0.359	−0.068	2.060
Standard error	0.058	0.046	0.131	0.144	0.155	0.123	0.137	0.145	0.161	0.122
Item 11										
Estimate	0.875	−0.991	0.318	−0.397	0.567	−0.116	0.613	0.501	0.834	0.817
Standard error	0.055	0.048	0.109	0.107	0.107	0.103	0.088	0.093	0.085	0.086
Item 12										
Estimate	2.087	0.992	−0.374	−1.895	−0.589	0.182	1.101	−1.492	0.542	1.008
Standard error	0.098	0.069	0.195	0.264	0.206	0.202	0.168	0.306	0.185	0.167
Item 13										
Estimate	1.626	−0.321	−0.718	0.558	−0.139	0.922	0.555	1.467	1.580	2.877
Standard error	0.078	0.055	0.216	0.211	0.201	0.196	0.226	0.197	0.197	0.185
Item 14										
Estimate	2.096	−0.804	−0.577	−0.696	−0.539	−0.221	−0.613	−1.659	−0.334	0.435
Standard error	0.102	0.067	0.114	0.158	0.107	0.092	0.097	0.147	0.089	0.070
Item 15										
Estimate	1.104	−1.803	−0.520	−0.781	−0.564	−0.680	−0.343	0.130	−0.730	−0.432
Standard error	0.067	0.065	0.087	0.078	0.096	0.089	0.067	0.061	0.075	0.070
Item 16										
Estimate	0.965	−1.060	−0.187	0.467	0.802	−0.199	1.074	0.209	0.407	−0.445
Standard error	0.057	0.050	0.092	0.113	0.112	0.125	0.076	0.086	0.084	0.106
Item 17										
Estimate	1.118	−1.309	0.310	0.512	1.364	0.149	2.761	3.379	1.632	1.423
Standard error	0.064	0.056	0.196	0.193	0.217	0.212	0.189	0.187	0.201	0.204
Item 18										
Estimate	1.781	−1.351	0.400	−0.291	0.321	0.097	1.451	0.316	1.619	2.397
Standard error	0.090	0.069	0.156	0.175	0.152	0.144	0.131	0.159	0.129	0.124
Item 19										
Estimate	0.675	−1.335	−0.936	−0.235	−0.812	−0.294	−1.112	0.342	−0.935	0.431
Standard error	0.053	0.050	0.110	0.074	0.104	0.073	0.103	0.061	0.094	0.060
Item 20										
Estimate	0.533	−1.463	−0.720	0.390	0.318	−0.680	−1.051	0.208	0.541	−0.578
Standard error	0.053	0.050	0.110	0.079	0.074	0.095	0.103	0.064	0.060	0.087

Table 5. Item parameters of the 3PNL model.

Item	Correct Response			Distractors							
	α_i	β_i	logit(γ_i)	$\lambda_{i,1}$	$\lambda_{i,2}$	$\lambda_{i,3}$	$\lambda_{i,4}$	$\delta_{i,1}$	$\delta_{i,2}$	$\delta_{i,3}$	$\delta_{i,4}$
Item 1											
Estimate	1.443	2.843	−3.065	0.396	0.927	0.668	1.132	−0.323	2.768	1.196	1.623
Standard error	0.125	0.231	4.501	0.474	0.304	0.342	0.343	0.761	0.500	0.552	0.532
Item 2											
Estimate	1.319	2.564	−4.277	−1.109	−0.271	−0.132	−0.452	−3.220	−1.141	−1.522	−1.709
Standard error	0.086	0.111	4.167	0.323	0.174	0.207	0.209	0.540	0.226	0.259	0.288

Table 5. *Cont.*

Item	Correct Response			Distractors							
	α_i	β_i	logit(γ_i)	$\lambda_{i,1}$	$\lambda_{i,2}$	$\lambda_{i,3}$	$\lambda_{i,4}$	$\delta_{i,1}$	$\delta_{i,2}$	$\delta_{i,3}$	$\delta_{i,4}$
Item 3											
Estimate	1.719	1.608	−2.921	−0.824	−0.424	−0.883	−0.496	−1.086	−1.130	−1.378	−0.549
Standard error	0.149	0.131	1.645	0.164	0.176	0.177	0.146	0.194	0.193	0.216	0.158
Item 4											
Estimate	0.625	1.181	−4.813	0.684	2.387	2.171	−0.076	2.602	6.763	3.888	−0.230
Standard error	0.051	0.065	4.023	0.588	0.581	0.606	0.769	1.136	1.100	1.115	1.535
Item 5											
Estimate	2.340	1.664	−3.412	0.536	0.127	0.300	−0.207	1.795	0.298	1.263	−0.421
Standard error	0.168	0.098	1.215	0.191	0.227	0.197	0.252	0.242	0.293	0.252	0.343
Item 6											
Estimate	1.363	1.418	−3.134	−1.265	−0.982	−1.262	−1.159	−2.399	−2.706	−2.295	−2.695
Standard error	0.125	0.158	2.550	0.183	0.217	0.176	0.208	0.230	0.264	0.220	0.263
Item 7											
Estimate	1.826	1.854	−6.090	−0.338	−0.431	−1.327	−1.119	−0.425	−0.565	−1.599	−1.751
Standard error	0.090	0.081	4.208	0.155	0.160	0.197	0.213	0.165	0.174	0.259	0.278
Item 8											
Estimate	1.378	1.394	−4.012	−0.752	−0.427	−0.002	−1.495	0.168	−0.543	0.512	−2.442
Standard error	0.103	0.123	4.313	0.147	0.176	0.141	0.263	0.154	0.183	0.133	0.381
Item 9											
Estimate	2.648	1.969	−2.061	0.408	0.533	0.422	0.297	1.305	0.828	0.176	1.664
Standard error	0.200	0.116	0.370	0.225	0.249	0.285	0.212	0.298	0.320	0.364	0.286
Item 10											
Estimate	1.461	−0.870	−1.983	0.317	0.456	0.250	1.137	0.701	0.356	−0.061	2.034
Standard error	0.152	0.172	0.274	0.119	0.132	0.141	0.113	0.133	0.141	0.156	0.118
Item 11											
Estimate	2.527	−3.084	−1.619	0.279	−0.374	0.581	−0.113	0.605	0.515	0.824	0.819
Standard error	0.315	0.385	0.096	0.106	0.102	0.106	0.099	0.087	0.092	0.085	0.085
Item 12											
Estimate	2.308	0.758	−2.504	−0.344	−1.748	−0.542	0.148	1.120	−1.412	0.573	0.990
Standard error	0.159	0.093	0.359	0.180	0.243	0.191	0.188	0.162	0.296	0.178	0.161
Item 13											
Estimate	1.593	−0.374	−7.481	−0.630	0.525	−0.116	0.852	0.615	1.446	1.596	2.837
Standard error	0.075	0.055	3.982	0.194	0.191	0.181	0.177	0.216	0.189	0.189	0.177
Item 14											
Estimate	2.249	−1.038	−3.833	−0.510	−0.646	−0.473	−0.183	−0.576	−1.635	−0.297	0.452
Standard error	0.142	0.102	0.434	0.104	0.144	0.098	0.085	0.094	0.143	0.086	0.069
Item 15											
Estimate	4.703	−6.146	−2.335	−0.596	−0.800	−0.590	−0.707	−0.344	0.144	−0.721	−0.422
Standard error	0.663	0.831	0.089	0.088	0.079	0.097	0.089	0.067	0.061	0.075	0.070
Item 16											
Estimate	4.626	−5.091	−1.638	−0.152	0.446	0.824	−0.214	1.089	0.205	0.404	−0.452
Standard error	0.608	0.675	0.072	0.088	0.112	0.115	0.118	0.075	0.086	0.084	0.105
Item 17											
Estimate	2.613	−3.013	−2.142	0.328	0.520	1.452	0.162	2.774	3.387	1.618	1.434
Standard error	0.277	0.313	0.117	0.182	0.180	0.211	0.198	0.188	0.186	0.201	0.202
Item 18											
Estimate	2.210	−1.798	−3.415	0.377	−0.242	0.321	0.122	1.444	0.342	1.618	2.407
Standard error	0.159	0.143	0.300	0.144	0.160	0.141	0.133	0.129	0.156	0.127	0.122
Item 19											
Estimate	3.167	−4.950	−1.672	−0.901	−0.241	−0.773	−0.300	−1.090	0.344	−0.911	0.434
Standard error	0.523	0.760	0.076	0.104	0.076	0.100	0.074	0.101	0.061	0.092	0.060
Item 20											
Estimate	2.233	−4.111	−1.827	−0.659	0.381	0.315	−0.628	−1.018	0.205	0.537	−0.551
Standard error	0.357	0.552	0.089	0.102	0.079	0.073	0.089	0.100	0.064	0.060	0.084

Table 6. Item parameters of the 4PNL model.

Item	Correct Response				Distractors							
	α_i	β_i	logit(γ_i)	logit(δ_i)	$\lambda_{i,1}$	$\lambda_{i,2}$	$\lambda_{i,3}$	$\lambda_{i,4}$	$\delta_{i,1}$	$\delta_{i,2}$	$\delta_{i,3}$	$\delta_{i,4}$
Item 1												
Estimate	2.447	3.234	0.395	0.983	0.413	0.955	0.677	1.174	−0.314	2.778	1.189	1.640
Item 2												
Estimate	1.733	2.875	0.149	0.983	−1.077	−0.283	−0.142	−0.448	−3.143	−1.150	−1.530	−1.697
Item 3												
Estimate	2.364	1.832	0.168	0.975	−0.801	−0.428	−0.904	−0.482	−1.052	−1.132	−1.392	−0.534
Item 4												
Estimate	1.517	2.702	0.016	0.852	0.690	2.391	2.175	0.049	2.619	6.790	3.913	0.015
Item 5												
Estimate	2.474	1.897	0.001	0.990	0.500	0.121	0.258	−0.309	1.749	0.288	1.212	−0.542
Item 6												
Estimate	2.063	1.698	0.192	0.956	−1.277	−1.048	−1.307	−1.155	−2.392	−2.766	−2.329	−2.673

Table 6. *Cont.*

Item	Correct Response				Distractors							
	α_i	β_i	logit(γ_i)	logit(δ_i)	$\lambda_{i,1}$	$\lambda_{i,2}$	$\lambda_{i,3}$	$\lambda_{i,4}$	$\delta_{i,1}$	$\delta_{i,2}$	$\delta_{i,3}$	$\delta_{i,4}$
Item 7												
Estimate	2.323	2.377	0.002	0.972	−0.336	−0.421	−1.341	−1.087	−0.424	−0.556	−1.613	−1.706
Item 8												
Estimate	2.260	1.575	0.225	0.955	−0.765	−0.426	0.005	−1.551	0.166	−0.539	0.515	−2.474
Item 9												
Estimate	3.508	2.443	0.159	0.985	0.447	0.560	0.460	0.325	1.343	0.851	0.212	1.691
Item 10												
Estimate	1.357	−0.757	0.104	0.999	0.332	0.473	0.278	1.152	0.711	0.368	−0.041	2.048
Item 11												
Estimate	2.444	−3.023	0.165	1.000	0.286	−0.378	0.583	−0.110	0.608	0.512	0.829	0.820
Item 12												
Estimate	2.766	0.948	0.098	0.976	−0.327	−1.817	−0.519	0.160	1.131	−1.481	0.590	0.997
Item 13												
Estimate	1.576	−0.356	0.000	1.000	−0.640	0.555	−0.096	0.890	0.607	1.466	1.611	2.861
Item 14												
Estimate	2.176	−0.980	0.019	1.000	−0.510	−0.661	−0.469	−0.177	−0.579	−1.646	−0.297	0.453
Item 15												
Estimate	4.743	−6.281	0.088	1.000	−0.585	−0.799	−0.588	−0.704	−0.348	0.136	−0.727	−0.428
Item 16												
Estimate	4.613	−5.115	0.162	1.000	−0.155	0.454	0.830	−0.225	1.087	0.210	0.413	−0.458
Item 17												
Estimate	2.496	−2.914	0.104	1.000	0.357	0.555	1.501	0.196	2.792	3.409	1.641	1.454
Item 18												
Estimate	2.146	−1.745	0.031	1.000	0.359	−0.264	0.303	0.102	1.437	0.332	1.611	2.398
Item 19												
Estimate	3.048	−4.844	0.157	0.998	−0.912	−0.245	−0.784	−0.300	−1.096	0.343	−0.917	0.433
Item 20												
Estimate	2.217	−4.113	0.139	0.991	−0.666	0.393	0.319	−0.633	−1.023	0.206	0.540	−0.555

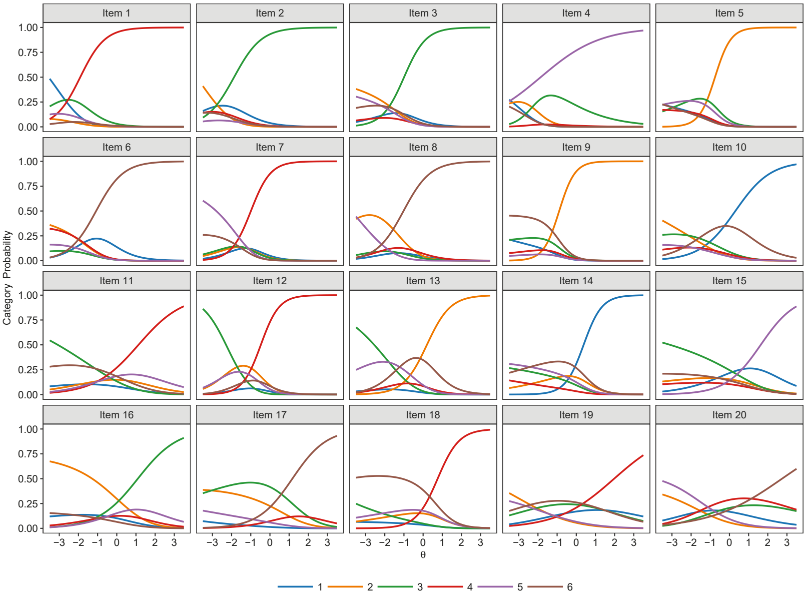


Figure 5. Item category curve plots of the 2-Parameter Nested Logit (2PNL) model.

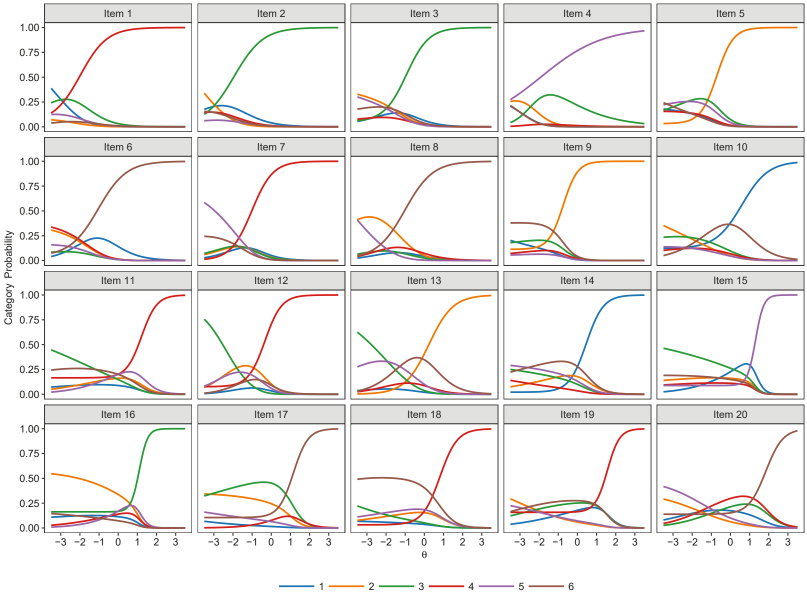


Figure 6. Item category curve plots of the 3-Parameter Nested Logit (3PNL) model.

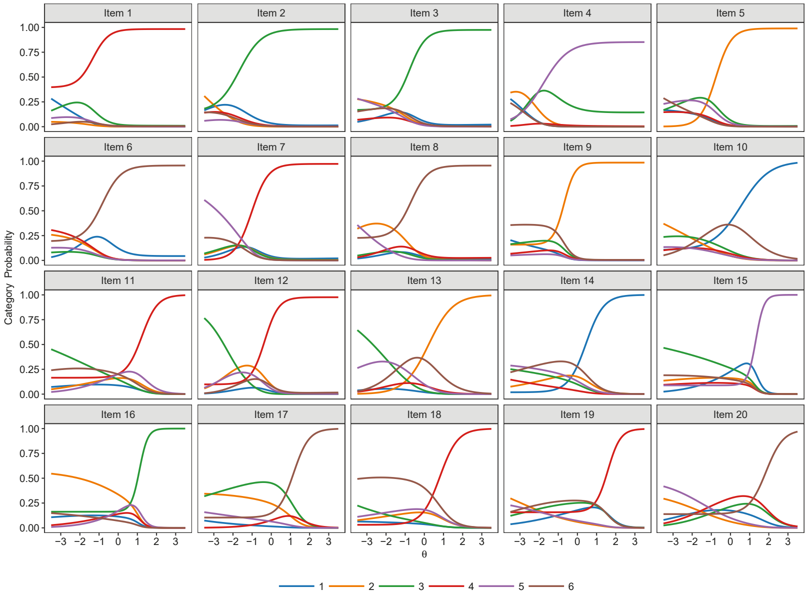


Figure 7. Item category curve plots of the 4-Parameter Nested Logit (4PNL) model.

The marginal estimates of empirical reliability for all the nominal models were satisfactory, as they were 0.857 for the Nominal Response model, 0.867 for the 2PNL model, 0.887 for the 3PNL model and 0.888 for the 4PNL model.

As hypothesized, preferring NLMs instead of binary logistic models resulted in significant reliability gains. The average reliability gains amounted to 0.018 (Bootstrapped 95% CI = [0.017, 0.021], Bootstrapped z = 17.765, $p < 0.001$) for the 2PL vs. 2PNL models, 0.019 (Bootstrapped 95% CI = [0.018, 0.023] , Bootstrapped z = 15.265, $p < 0.001$) for the 3PL vs. 3PNL models, and 0.015 (Bootstrapped 95% CI = [0.011, 0.020] , Bootstrapped z = 6.669, $p < 0.001$) for the 4PL vs. 4PNL models.

The reliability functions of the 2PL, 3PL and 4PL are reported with their Nested Logit counterparts in respectively Figures 8–10. As noted by a reviewer, between a binary model and its nested counterpart, θ_j is not perfectly invariant, and thus the reliability functions may cross, such as in Figure 4. This was also previously observed in the comparison between binary and nominal response models [7].

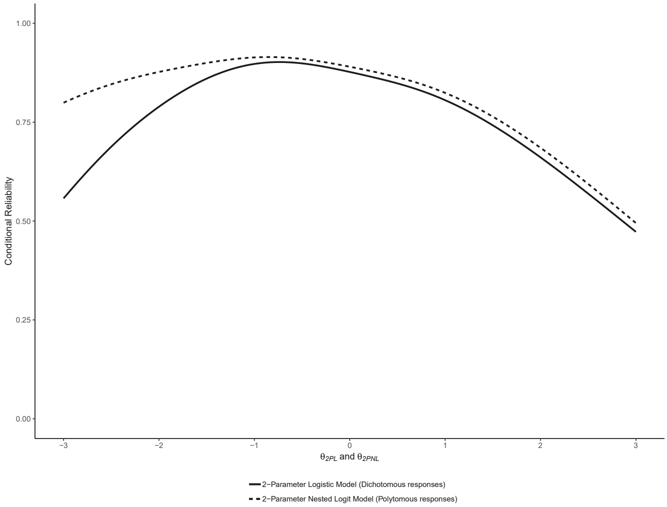


Figure 8. Comparison of the reliability functions of the 2-Parameter Logistic (2PL) and Nested Logit (2PNL) models.

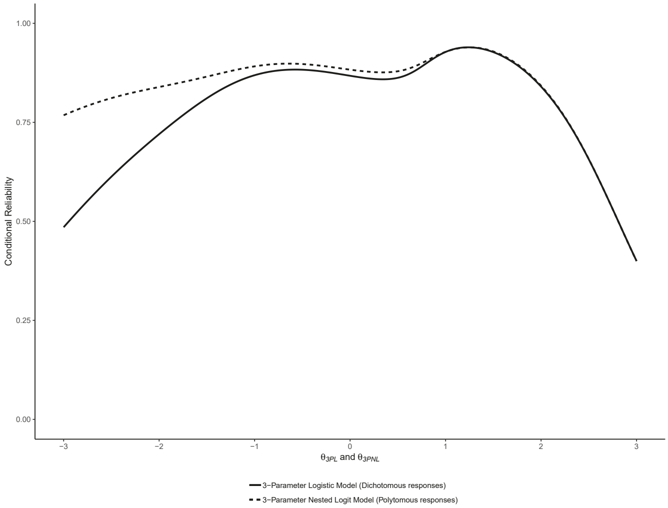


Figure 9. Comparison of the reliability functions of the 3-Parameter Logistic (3PL) and Nested Logit (3PNL) models.

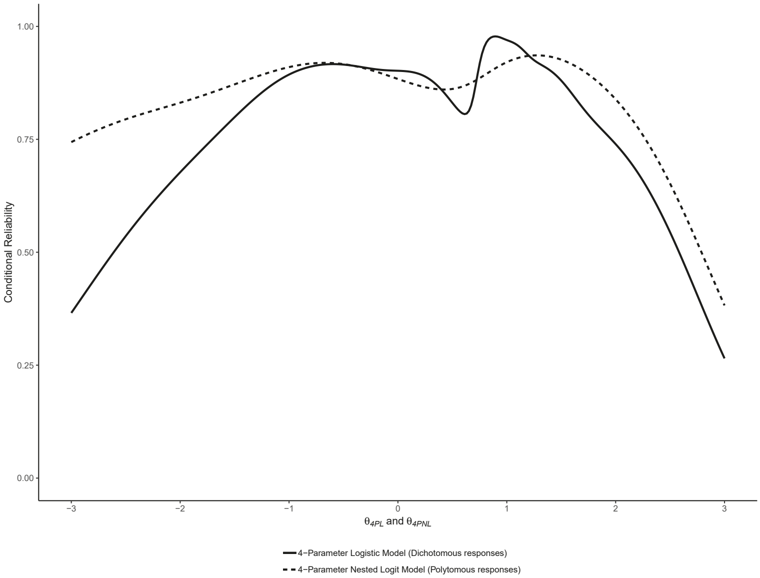


Figure 10. Comparison of the reliability functions of the 4-Parameter Logistic (4PL) and Nested Logit (4PNL) models.

As expected, they show that using NLM provided increments in reliability especially in the lower range of abilities.

4. Discussion

The aim of the present research was to extend the previous findings of Myszkowski and Storme [8] to different testing modalities, online assessment—a different context with higher stakes—and personnel selection, on a larger sample and with a different logical reasoning test.

We found that 4-parameter models—both binary and nested logit—were likely unstable (as their information matrix could not be inverted) but they seemed to outperform their 1PL, 2PL, and 3PL counterparts. Being that the 2-parameter and 3-parameter models did not present this issue while still presenting excellent fit, the results suggest that choosing them may be a more parsimonious but still well fitting approach to this test. In fact, the 2PL and 2PNL fitting respectively almost as well as the 3PL and 3PNL, they may be a more optimal modeling strategy for this test.

We also found that, as hypothesized, Nested Logit Models (NLM) both outperformed the Nominal Response Model [7], providing significant reliability gains compared with their binary counterparts. In addition, the absolute fit of the NLMs—which was not computable in Myszkowski and Storme [8] due to the lower sample size—could be computed here and was found satisfactory, especially regarding the models including a guessing parameter (3PNL and 4PNL).

These findings overall suggest that NLMs [9] are a better modeling alternative than binary logistic models and than the Nominal Response Model [7] for logical reasoning multiple-choice tests, such as incomplete matrix or series tests, in online personnel selection settings.

4.1. Theoretical and Practical Implications

From a theoretical viewpoint, the present study can be seen as a conceptual replication and extension of Myszkowski and Storme [8]'s study on Raven's progressive matrices. Replicating findings is an important endeavor in scientific research. This is especially true in the field of psychology, which is regularly criticized for its lack of consideration for replicating empirical findings [27].

Recently, Hüffmeier et al. [10] have designed a theoretical framework to conceptualize the replication process in psychology and have proposed a typology of replication studies. Rather than considering replication as a process separate from the initial research process, they conceptualize replication as the very research process by which fundamental findings are generalized to situations that are increasingly close to real life conditions.

When a result has been shown at a fundamental level, it may be interesting to replicate it to see if it is not due to chance. In this case, exact or close replications will be used [10]. To be able to further generalize the findings of a fundamental study, it is important to be able to perform conceptual replications in the laboratory or in the field. In conceptual replications, comparability to the original study is limited to the aspects that are considered theoretically relevant [28,29]. Among the conceptual replications are field studies. The aim of such studies is to investigate whether laboratory findings also hold under field conditions, and to rule out the possibility that a laboratory finding is a laboratory artifact or too weak to be relevant in contexts that are less tightly controlled [10]. In the framework described by Hüffmeier et al. [10], our study can be defined as a conceptual replication in the field of the study conducted by Myszkowski and Storme [8]. Our findings suggest that the characteristics of the e-assessment context do not fundamentally affect the way distractors are selected by test takers. Previous basic research on recovering distractor information is therefore relevant in an e-assessment context.

From a practical viewpoint, our findings suggest that one way to improve the accuracy of e-assessment in the context of recruitment is to recover distractor information. Web applications that use tests with distractors should try to implement NLM to get more reliable estimates of the general mental ability of job applicants. To this day, there are few software implementations of NLM. A recommendation to designers of IRT platforms would be to add NLM to their offer. For e-assessment platforms, a relatively inexpensive alternative to commercial IRT software could be to use the "mirt" [23] R library on the server side to estimate the ability of test takers using the built-in NLM function. One of the challenges of this option is that R can be a programming language that is relatively consuming in terms of computing resources and time, although θ_j estimations in "mirt" are relatively fast once the parameters of the model are stored in memory. More optimizations that will facilitate the implementation of NLM in e-assessment might come in the future.

In line with the findings of Myszkowski and Storme [8], the observed gain in reliability was especially visible at relatively low levels of ability. This is not surprising as NLM recover information from wrong response options. Recruiters are usually interested in applicants with high levels of intelligence, but this is not always the case. For example, it is possible that due to high competition on the job market, a recruiter is unable to attract the best applicants, and has to select among applicants with relatively lower levels of ability. In such situations, the use of NLM could be highly valuable as it allows forming a more accurate impression of applicants on the low end of the trait, and selecting the best.

As a reviewer pointed out, the standard errors of item parameter estimates of the Nested Logit Models were overall smaller than their binary counterparts—this of course only concerns parameters that are common between models (difficulty, discrimination and, for the 3PL and 3PNL, guessing). This result may seem counterintuitive, because, in general, for a given dataset, item parameter standard errors tend to increase as model complexity increases, and the Nested Logit Models are substantially more parametrized than the binary models. However, it should be noted that the Nested Logit Models are not only more complex, but they also use, to some extent, a different dataset, in that they use more information from the base dataset. Indeed, they use the complete information from the nominal level, while binary models use only the information at the binary level. Although we have showed that, like in Myszkowski and Storme [8], Nested Logit Models resulted in gains in reliability (and thus lower standard errors) for the person estimates, the present results also suggest that the difficulty, disscrimination and guessing parameters of the Nested Logit Models are estimated with more accuracy—because they use more information—than the respective item parameters of their binary counterparts. This result calls for replication in other datasets, contexts and types of tests.

Throughout the paper, we have mostly emphasized the benefits of using NLM to improve the accuracy of ability estimates. However, NLM has other potentially interesting applications beyond improving scoring. For example, Suh and Bolt [30] have described a method relying on NLM to evaluate how distractors might contribute to Differential Item Functioning (DIF) [30]. It is indeed possible that distractors function differently across groups, leading to Differential Distractor Functioning (DDF). DDF can lead in turn to DIF, which is a major problem when using the same test on different groups. Multigroup NLM could help test designers to improve the diagnosis of the causes of DIF, and thus to improve their tests. Bolt et al. [31] have suggested another interesting application of NLM, which is to use NLM as a way to determine whether the ability distinguished by distractors is the same as the ability underlying the choice of the correct response. Here again, the use of NLM could help test designers to select items that best reflect the underlying ability.

4.2. Limitations and Future Research

Our study has several limitations which should stimulate and guide further research on the topic. A first limitation is related to the sample that was used in the study. The sample comes from a single e-assessment platform and it is therefore difficult to know whether the findings would generalize to other platforms. It is possible for example that characteristics of the design of Web applications affect the way distractors are processed by test takers. Previous research has shown that the experience of users greatly affect the cognitive processes they mobilize when using a Web application [32]. Applied to our question, it is possible that a bad Web design reduces the motivation of test takers to process distractors when they fail at identifying the rule governing the logical progression of the series. Further research is needed to test the generalizability of the findings to other platforms, but also to other types of GMA tasks.

Antother limitation is related to our sample size. NLM have more parameters than the models to which they were compared in the current study. Although our sample is larger than the one used in the original study that we conceptually replicated [8], it is still unclear whether our sample size is large enough to get reliable parameter estimates. Further research using Monte-Carlo simulations is needed on the influence of sample size on parameter estimation in NLM, and to provide clear guidelines regarding the necessary sample size.

In addition, it should be noted that the fact that NLM provided a better fit, like in Myszkowski and Storme [8]'s study, does not necessarily imply that the cognitive processes engaged in responding similar tests are necessarily only the 2-step sequence that the NLM are based on—attempting to solve the task by looking at the stimulus only and then, if the correct answer is not found, examining the distractors. Indeed, it remains very possible that the actual responding process is less clear and closer to a back-and-forth between a stimulus-based strategy and a response option comparison-based strategy. Further, it has been noted that NLM may be further improved by including the possibility that the guessing strategy (level 2) results in the choice of the correct response. In other words, choosing the correct response could then be the result or either strategy. Future research might consider this interesting possibility when such models are available in traditional IRT software.

Another limitation of this study is that it was limited in the breadth of nominal models tested by their availibility in "mirt." Although this package provides a large number of popular models, we were not able to fit some alternatives models, notably Thissen and Steinberg [33]'s Multiple Choice Model (MCM), which essentially adds to the Nominal Response model a latent state category for examinees that corresponds to an examinee not knowing—and thus guessing—what the correct response is. Although the Nominal Response model was here outperformed by the Nested Logit models, it may be that alternative models like the MCM are better alternatives.

Another important limitation of our study is that we did not test whether the improvement in reliability translates into an improvement in predictive validity. This is because our study did not include a measure of job performance. The ability of an assessment tool to predict future job performance is crucial in the context of recruitment. Improvements in measurement reliability can lead to improvements in predictive validity, as reliability is a prerequisite for validity [34].

Whether recovering distractor information actually improves predictive validity in the context of e-assessment remains to be investigated. The answer to this question could represent an important contribution to the literature. It has indeed been shown that in situations in which test takers are under pressure, for example when stakes are high, the predictive validity of GMA tests tends to decrease [35]. Duckworth et al. [35] argued that GMA tests predict various indicators of success in life because when they are used in low stakes contexts, they essentially measure the motivation of test takers. According to Duckworth et al. [35], it is because GMA tests taken in the lab measure motivation that they are found to be positively associated with a broad range of indicators of life success. Although there is empirical evidence supporting Duckworth et al. [35]'s argument, one can wonder whether using a more precise strategy to score GMA tests could not ultimately reveal that there is a relation between GMA and various indicators of achievement. Testing the predictive validity of GMA tests scored with NLM could therefore have important implications regarding the knowledge of the true relationship between GMA and achievement in general.

Author Contributions: Conceptualization, M.S. and N.M.; methodology, M.S., S.B. and D.B.; investigation, M.S.; data curation, S.B. and D.B.; Writing—Original Draft preparation, N.M. and M.S.; Writing—Review and Editing, N.M. and M.S.; visualization, M.S.; supervision, M.S.; project administration, M.S., N.M., S.B. and D.B.

Funding: This research received no external funding.

Conflicts of Interest: Authors 3 and 4 hold positions in the company that owns the psychometric test used (GF20) in this study.

References

1. Bartram, D. Internet recruitment and selection: Kissing frogs to find princes. *Int. J. Sel. Assess.* **2000**, *8*, 261–274. [CrossRef]
2. Laumer, S.; von Stetten, A.; Eckhardt, A. E-assessment. *Bus. Inf. Syst. Eng.* **2009**, *1*, 263–265. [CrossRef]
3. Schmidt, F.L.; Hunter, J. General mental ability in the world of work: Occupational attainment and job performance. *J. Personal. Soc. Psychol.* **2004**, *86*, 162. [CrossRef] [PubMed]
4. Ryan, A.M.; Ployhart, R.E. Applicants' perceptions of selection procedures and decisions: A critical review and agenda for the future. *J. Manag.* **2000**, *26*, 565–606. [CrossRef]
5. Gilliland, S.W.; Steiner, D.D. Causes and consequences of applicant perceptions of unfairness. In *Justice in the Workplace*; Cropanzano, R., Ed.; Erlbaum: Hillsdale, NJ, USA, 2001; pp. 175–195.
6. Tavakol, M.; Dennick, R. Making sense of Cronbach's alpha. *Int. J. Med. Educ.* **2011**, *2*, 53. [CrossRef] [PubMed]
7. Bock, R.D. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika* **1972**, *37*, 29–51. [CrossRef]
8. Myszkowski, N.; Storme, M. A snapshot of g? Binary and polytomous item-response theory investigations of the last series of the Standard Progressive Matrices (SPM-LS). *Intelligence* **2018**, *68*, 109–116. [CrossRef]
9. Suh, Y.; Bolt, D.M. Nested logit models for multiple-choice item response data. *Psychometrika* **2010**, *75*, 454–473. [CrossRef]
10. Hüffmeier, J.; Mazei, J.; Schultze, T. Reconceptualizing replication as a sequence of different studies: A replication typology. *J. Exp. Soc. Psychol.* **2016**, *66*, 81–92. [CrossRef]
11. Edelen, M.O.; Reeve, B.B. Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Qual. Life Res.* **2007**, *16*, 5. [CrossRef]
12. Kim, S.; Feldt, L.S. The estimation of the IRT reliability coefficient and its lower and upper bounds, with comparisons to CTT reliability statistics. *Asia Pac. Educ. Rev.* **2010**, *11*, 179–188. [CrossRef]
13. Hambleton, R.K.; Van der Linden, W.J. Advances in item response theory and applications: An introduction. *Appl. Psychol. Meas.* **1982**, *6*, 373–378. [CrossRef]
14. Yen, Y.C.; Ho, R.G.; Laio, W.W.; Chen, L.J.; Kuo, C.C. An empirical evaluation of the slip correction in the four parameter logistic models with computerized adaptive testing. *Appl. Psychol. Meas.* **2012**, *36*, 75–87. [CrossRef]

15. Myszkowski, N.; Storme, M. Measuring "good taste" with the visual aesthetic sensitivity test-revised (VAST-R). *Personal. Individ. Diff.* **2017**, *117*, 91–100. [CrossRef]
16. Martín, E.S.; del Pino, G.; Boeck, P.D. IRT Models for Ability-Based Guessing. *Appl. Psychol. Meas.* **2006**, *30*, 183–203. [CrossRef]
17. Matzen, L.B.V.; Van der Molen, M.W.; Dudink, A.C. Error analysis of Raven test performance. *Personal. Individ. Diff.* **1994**, *16*, 433–445. [CrossRef]
18. Raven, J.C. Standardization of progressive matrices, 1938. *Br. J. Med. Psychol.* **1941**, *19*, 137–150. [CrossRef]
19. Beilock, S.L.; Carr, T.H. When high-powered people fail: Working memory and "choking under pressure" in math. *Psychol. Sci.* **2005**, *16*, 101–105. [CrossRef]
20. Gimmig, D.; Huguet, P.; Caverni, J.P.; Cury, F. Choking under pressure and working memory capacity: When performance pressure reduces fluid intelligence. *Psychon. Bull. Rev.* **2006**, *13*, 1005–1010. [CrossRef]
21. Jorgensen, T.D.; Pornprasertmanit, S.; Miller, P.; Schoemann, A.; Rosseel, Y.; Quick, C.; Garnier-Villarreal, M.; Selig, J.; Boulton, A.; Preacher, K.; et al. semTools: Useful Tools for Structural Equation Modeling. Available online: https://cran.r-project.org/web/packages/semTools/semTools.pdf (accessed on 10 July 2019).
22. Rosseel, Y. Lavaan: An R Package for Structural Equation Modeling. *J. Stat. Softw.* **2012**, *48*. [CrossRef]
23. Chalmers, R.P. mirt: A multidimensional item response theory package for the R environment. *J. Stat. Softw.* **2012**, *48*, 1–29. [CrossRef]
24. Myszkowski, N.; Storme, M. Judge Response Theory? A Call to Upgrade Our Psychometrical Account of Creativity Judgments. *Psychol. Aesthet. Creat. Arts* **2019**, *13*, 167–175. [CrossRef]
25. Hansen, M.; Cai, L.; Monroe, S.; Li, Z. Limited-Information Goodness-of-Fit Testing of Diagnostic Classification Item Response Theory Models. CRESST Report 840. *Natl. Center Res. Eval. Stand. Stud. Test. (CRESST)* **2014**, *1*, 1–47.
26. Raju, N.S.; Price, L.R.; Oshima, T.; Nering, M.L. Standardized conditional SEM: A case for conditional reliability. *Appl. Psychol. Meas.* **2007**, *31*, 169–180. [CrossRef]
27. Fabrigar, L.R.; Wegener, D.T. Conceptualizing and evaluating the replication of research results. *J. Exp. Soc. Psychol.* **2016**, *66*, 68–80. [CrossRef]
28. Schmidt, S. Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Rev. Gen. Psychol.* **2009**, *13*, 90–100. [CrossRef]
29. Stroebe, W.; Strack, F. The alleged crisis and the illusion of exact replication. *Perspect. Psychol. Sci.* **2014**, *9*, 59–71. [CrossRef]
30. Suh, Y.; Bolt, D.M. A nested logit approach for investigating distractors as causes of differential item functioning. *J. Educ. Meas.* **2011**, *48*, 188–205. [CrossRef]
31. Bolt, D.M.; Wollack, J.A.; Suh, Y. Application of a multidimensional nested logit model to multiple-choice test items. *Psychometrika* **2012**, *77*, 339–357. [CrossRef]
32. Abbey, B. *Instructional and Cognitive Impacts of Web-Based Education*; IGI Global: Dauphin County, PA, USA, 1999.
33. Thissen, D.; Steinberg, L. A response model for multiple choice items. *Psychometrika* **1984**, *49*, 501–519. [CrossRef]
34. Davidshofer, K.; Murphy, C.O. *Psychological Testing: Principles and Applications*; Pearson/Prentice HallUpper: Saddle River, NJ, USA, 2005.
35. Duckworth, A.L.; Quinn, P.D.; Lynam, D.R.; Loeber, R.; Stouthamer-Loeber, M. Role of test motivation in intelligence testing. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 7716–7720. [CrossRef] [PubMed]

Journal of
Intelligence

MDPI

Article

Analysing Standard Progressive Matrices (SPM-LS) with Bayesian Item Response Models

Paul-Christian Bürkner

Department of Computer Science, Aalto University, Konemiehentie 2, 02150 Espoo, Finland;
paul.buerkner@gmail.com

Received: 21 November 2019; Accepted: 29 January 2020; Published: 4 February 2020

Abstract: Raven's Standard Progressive Matrices (SPM) test and related matrix-based tests are widely applied measures of cognitive ability. Using Bayesian Item Response Theory (IRT) models, I reanalyzed data of an SPM short form proposed by Myszkowski and Storme (2018) and, at the same time, illustrate the application of these models. Results indicate that a three-parameter logistic (3PL) model is sufficient to describe participants dichotomous responses (correct vs. incorrect) while persons' ability parameters are quite robust across IRT models of varying complexity. These conclusions are in line with the original results of Myszkowski and Storme (2018). Using Bayesian as opposed to frequentist IRT models offered advantages in the estimation of more complex (i.e., 3–4PL) IRT models and provided more sensible and robust uncertainty estimates.

Keywords: Standard Progressive Matrices; Item Response Theory; Bayesian statistics; brms; Stan; R

1. Introduction

Raven's Standard Progressive Matrices (SPM) test (Raven 1941) matrix-based tests are widely applied measures of cognitive ability (e.g., Jensen et al. 1988; Pind et al. 2003). Due to their non-verbal content, which reduces biases due to language and cultural differences, they are considered one of the purest measures of fluid intelligence (Myszkowski and Storme 2018). However, a disadvantage of the original SPM is that its administration takes considerable time as 60 items have to be answered and time limits are either very loose or not imposed at all (e.g., Pind et al. 2003). Thus, using it as part of a bigger procedure involving the administration of the SPM as part of a battery of tests and/or experiments may be problematic. This is not only due to direct time restrictions but also because participants' motivation and concentration tends to decline over the course of the complete procedure, potentially leading to less valid measurements (e.g., Ackerman and Kanfer 2009).

Recently, Myszkowski and Storme (Myszkowski and Storme 2018) proposed a short version of the original SPM test, called SPM-LS, comprising only the last block of the 12 most complex SPM items. They evaluated the statistical properties of the SPM-LS using methods of Item Response Theory (IRT). IRT is widely applied in the human sciences to model persons' responses on a set of items measuring one or more latent constructs (for a comprehensive introduction see Embretson and Reise 2013; Lord 2012; van der Linden and Hambleton 1997). Due to its flexibility compared to Classical Test Theory (CTT), IRT provides the formal statistical basis for most modern psychological measurement. The best known IRT models are likely those for binary responses, which predict the probability of a correct answer depending on item's properties and the participant's latent abilities. As responses on SPM items can be categorized as either right or wrong, I focus on these binary models in the present paper (although other models for these data are possible as well; see Myszkowski and Storme 2018). Myszkowski and Storme (Myszkowski and Storme 2018), whose data I sought to reanalyze, used frequenstist IRT models for inference. In this paper, I apply Bayesian IRT models instead and investigate potential differences to the original results. In doing so, I hope to improve our understanding of the robustness of the inference obtainable from the SPM-LS test and to illustrate the application of Bayesian IRT methods.

2. Bayesian IRT Models

In Bayesian statistics applied to IRT, we aim to estimate the posterior distribution $p(\theta, \xi|y)$ of the person and item parameters (θ and ξ, respectively, which may vary in number depending on the model) given the data y. We may be either interested in the posterior distribution directly, or in quantities that can be computed on its basis. The posterior distribution for an IRT model is defined as

$$p(\theta, \xi|y) = \frac{p(y|\theta, \xi)\, p(\theta, \xi)}{p(y)}. \tag{1}$$

In the above equation, $p(y|\theta, \xi)$ is the likelihood, $p(\theta, \xi)$ is the prior distribution, and $p(y)$ is the marginal likelihood. The likelihood $p(y|\theta, \xi)$ is the distribution of the data given the parameters and thus relates the data to the parameters. tThe prior distribution $p(\theta, \xi)$ describes the uncertainty in the person and item parameters before having seen the data. It thus allows explicitly incorporating prior knowledge into the model and/or helping to identify the model. In practice, we factorize the joint prior $p(\theta, \xi)$ into the product of $p(\theta)$ and $p(\xi)$ so that we can specify priors on person and items parameters independently. I provide more details on likelihoods and priors for Bayesian IRT models in the next section. The marginal likelihood $p(y)$ serves as a normalizing constant so that the posterior is an actual probability distribution. Except in the context of specific methods (e.g., Bayes factors), $p(y)$ is rarely of direct interest.

Obtaining the posterior distribution analytically is only possible in certain cases of carefully chosen combinations of prior and likelihood, which may considerably limit modelling flexibility but yield a computational advantage. However, with the increased power of today's computers, Markov-Chain Monte-Carlo (MCMC) sampling methods constitute a powerful and feasible alternative to obtaining posterior distributions for complex models in which the majority of modeling decisions is made based on theoretical and not computational grounds. Despite all the computing power, these sampling algorithms are computationally very intensive and thus fitting models using full Bayesian inference is usually much slower than in point estimation techniques. If using MCMC to fit a Bayesian model turns out to be infeasible, an alternative is to perform optimization over the posterior distribution to obtain Maximum A-Posteriori (MAP) estimates, a procedure similar to maximum likelihood estimation just with additional regularization through priors. MCMC and MAP estimates differ in at least two aspects. First, MCMC allows obtaining point estimates (e.g., means or medians) from the unidimensional marginal posteriors of the quantities of interest, which tend to be more stable than MAP estimates obtained from the multidimensional posterior over all parameters. Second, in contrast to MAP, MCMC provides a set of random draws from the model parameters' posterior distribution. After the model fitting, the posterior distribution of any quantity that is a function of the original parameters can be obtained by applying the function on a draw by draw basis. As such, the uncertainty in the posterior distribution naturally propagates to new quantities, a highly desirable property that is difficult to achieve using point estimates alone.

In the present paper, I apply Bayesian binary IRT models to the SPM-LS data using both MCMC and MAP estimators. Their results are compared to those obtained by frequentist maximum likelihood estimation. For a comprehensive introduction to Bayesian IRT modeling see, for example, the works of Fox (Fox 2010), Levy and Mislevy (Levy and Mislevy 2017), and Rupp, Dey, and Zumbo (Rupp et al. 2004).

2.1. Bayesian IRT Models for Binary Data

In this section, I introduce a set of Bayesian IRT models for binary data and unidimensional person traits. Suppose that, for each person j ($j = 1, \ldots, J$) and item i ($i = 1, \ldots, I$), we have observed a binary response y_{ji}, which is coded as 1 for a correct answer and 0 otherwise. With binary IRT models, we aim to model $p_{ji} = P(y_{ji} = 1)$, that is, the probability the person j answers item i correctly. In other words, we assume a Bernoulli distribution for the responses y_{ji} with success probability p_{ji}:

$$y_{ji} \sim \mathrm{Bernoulli}(p_{ji}) \tag{2}$$

Across all models considered here, we assume that all items measure a single latent person trait θ_j. For the present data, we can expect θ_j to represent something closely related to fluid intelligence (Myszkowski and Storme 2018). The most complex model I consider in this paper is the four-parameter logistic (4PL) model and all other simpler models result from this model by fixing some item parameters to certain values. In recent years, the 4PL model has received much attention in IRT research due to its flexibility in modeling complex binary response processes (e.g., Culpepper 2016, 2017; Loken and Rulison 2010; Waller and Feuerstahler 2017). Under this model, we express $P(y_{ji} = 1)$ via the equation

$$P(y_{ji} = 1) = \gamma_i + (1 - \gamma_i - \psi_i) \frac{1}{1 + \exp(-(\beta_i + \alpha_i \theta_j))}. \tag{3}$$

In the 4PL model, each item has four associated item parameters. The β_i parameter describes the location of the item, that is, how easy or difficult it is in general. In the above formulation of the model, higher values of β_i imply higher success probabilities and hence β_i can also be called the "easiness" parameter. The α_i parameter describes how strongly item i is related to the latent person trait θ_j. We can call α_i "factor loading", "slope", or "discrimination" parameter, but care must be taken that none of these terms is used uniquely and their exact meaning can only be inferred in the context of a specific model (e.g., see Bürkner 2019 for a somewhat different use of the term "discrimination" in IRT models). For our purposes, we assume α_i to be positive as we expect answering the items correctly implies higher trait scores than when answering incorrectly. In addition, if we did not fix the sign of α_i, we may run into identification issues as changing the sign of α_i could be compensated by changing the sign of θ_j without a change in the likelihood.

The γ_i parameter describes the guessing probability, that is, the probability of any person answering item i correctly even if they do not know the right answer and thus have to guess. For obvious reasons, guessing is only relevant if the answer space is reasonably small. In the present data, participants saw a set of 8 possible answers of which exactly one was considered correct. Thus, guessing cannot be ruled out and would be equal to $\gamma_i = 1/8$ for each item if all answer alternatives had a uniform probability to be chosen given that a person guesses. Lastly, the ψ_i parameter enables us to model the possibility that a participant makes a mistake even though they know the right answer, perhaps because of inattention or simply misclicking when selecting the chosen answer. We may call ψ_i the "lapse", "inattention", or "slipping" parameter. Usually, these terms can be used interchangeably but, as always, the exact meaning can only be inferred in the context of the specific model. As the answer format in the present data (i.e., "click on the right answer") is rather simple and participants have unlimited time for each item, mistakes due to lapses are unlikely to appear. However, by including a lapse parameter into our model, we are able to explicitly check whether lapses played a substantial role in the answers.

We can now simplify the 4PL model in several steps to yield the other less complex models. The 3PL model results from the 4PL model by additionally fixing the lapse probability to zero, that is, $\psi_i = 0$ for all items. In the next step, we can obtain the 2PL model from the 3PL model by also fixing the guessing probabilities to zero, that is, $\gamma_i = 0$ for all items. In the last simplification step, we obtain the 1PL model (also known as Rasch model Rasch 1961) from the 2PL model by assuming the factor loadings to be one, that is, $\alpha_i = 1$ for all items. Even though didactically I find it most intuitive and helpful to introduce the models from most to least complex, I recommend the inverse order in applications, that is, starting from the simplest (but still sensible) model. The reason is that more complex models tend to be more complicated to fit in the sense that they both take longer (especially when using MCMC estimation) and yield more convergence problems (e.g., Bürkner 2019; Gelman et al. 2013). If we started by fitting the most complex model and, after considerable waiting time, found the model to not have converged, we may have no idea which of the several model components were causing the problem(s). In contrast, by starting simple and gradually building towards more complex models, we can make sure that each model component is reasonably specified and can be reliably estimated before we move further. As a result, when a problem occurs, we are likely to have much clearer understanding of why/where it occurred and how to fix it.

With the model likelihood fully specified by Equations (2) and (3) (potentially with some fixed item parameters), we are, in theory, already able to obtain estimates of person and item parameters via maximum likelihood (ML) estimation. However, there are multiple potential issues that can get into our way at this point. First, we simply may not have enough data to obtain sensible parameter estimates. As a rule of thumb, the more complex a model, the more data we need to obtain the same estimation precision. Second, there may be components in the model which will not be identified no matter how much data we add. An example would be binary IRT models from 2PL upwards as (without additional structure) we cannot identify the scale of both θ_j and α_i at the same time. This is because, due to the multiplicative relationship, multiplying one of the two by a constant can be adjusted for by dividing the other by the same constant without changing the likelihood. Third, we need to have software that is able to do the model fitting for us, unless we want to hand code every estimation algorithm on our own. Using existing software requires (re)expressing our models in a way the software understands. I will focus on the last issue first and then address the former two.

2.2. IRT Models as Regression Models

There are a lot of IRT specific software packages available, in particular in the programming language R (R Core Team 2019), for example, mirt (Chalmers 2012), sirt (Robitzsch 2019), or TAM (Robitzsch et al. 2019; see Bürkner 2019 for a detailed comparison). In addition to these more specialized packages, general purpose probabilistic programming languages can be used to specify and fit Bayesian IRT models, for example, BUGS (Lunn et al. 2009; see also Curtis 2010), JAGS (Plummer 2013; see also Depaoli et al. 2016; Zhan et al. 2019), or Stan (Carpenter et al. 2017; see also Allison and Au 2018; Luo and Jiao 2018). In this paper, I use the brms package (Bürkner 2017, 2018), a higher level interface to Stan, which is not focused specifically on IRT models but more generally on (Bayesian) regression models. Accordingly, we need to rewrite our IRT models in a form that is understandable for brms or other packages focussed on regression models.

The first implication of this change of frameworks is that we now think of the data in long format, with all responses from all participants on all items in the same data column coupled with additional columns for person and item indicators. That is, y_{ji} is now formally written as y_n where n is the observation number ranging from 1 to $N = J \times I$. If we needed to be more explicit we could also use $y_{j_n i_n}$ to indicate that each observation number n has specific indicators j and i associated with it. The same goes for item and person parameters. For example, we may write θ_{n_j} to refer to the ability parameter of the person j to whom the nth observation belongs.

One key aspect of regression models is that we try to express parameters on an unconstrained space that spans the whole real line. This allows for using linear (or more generally additive) predictor terms without having to worry about whether these predictor terms fulfill certain boundaries, for instance, are positive or within the unit interval $[0, 1]$. In the considered binary IRT models, we need to ensure that the factor loadings α are positive and that guessing and lapse parameters, γ and ψ, respectively, are within $[0, 1]$ as otherwise the interpretation of the latter two as probabilities would not be sensible. To enforce these parameter boundaries within a regression, we apply (inverse-)link functions. That is, for α, we use the log-link function (or equivalently the exponential response function) so that

$$\alpha = \exp(\eta_\alpha) \tag{4}$$

where η_{α_n} is unconstrained. Similarly, for γ and ψ, we use the logit-link (or equivalently the logistic response function) so that

$$\gamma = \text{logistic}(\eta_\gamma) = \frac{1}{1 + \exp(-\eta_\gamma)}, \tag{5}$$

$$\psi = \text{logistic}(\eta_\psi) = \frac{1}{1 + \exp(-\eta_\psi)} \tag{6}$$

where η_γ and η_ψ are unconstrained. The location parameters β are already unbounded and as such do not need an additional link function so that simply $\beta = \eta_\beta$. The same goes for the ability parameters θ. On the scale of the linear predictors, we can perform the usual regression operations, perhaps most importantly modeling predictor variables or including multilevel structure. In the present data, we do not have any additional person or item variables available so there are no such predictors in our models (but see Bürkner 2019 for examples if you are interested in this option). However, there certainly is multilevel structure as we have both multiple observations per item and per person, which we seek to model appropriately, as detailed in the next section.

2.3. Model Priors and Identification

When it comes to the specification of priors on item parameters, we typically distinguish between non-hierarchical and hierarchical priors (Bürkner 2019; Fox 2010; Levy and Mislevy 2017) with the former being applied more commonly (e.g., Bürkner 2018; Levy and Mislevy 2017). When applying non-hierarchical priors, we directly equate the linear predictor η (for any of the item parameter classes) with item-specific parameters b_i, so that

$$\eta_n = b_{i_n} \tag{7}$$

for each observation n and corresponding item i. Since η is on an unconstrained scale so are the b_i parameters and we can apply location-scale priors such as the normal distribution with mean μ and standard deviation σ:

$$b_i \sim \text{normal}(\mu, \sigma) \tag{8}$$

In non-hierarchical priors, we fix μ and σ to sensible values. In general, priors can only be understood in the context of the model as a whole, which renders general recommendation for prior specification difficult (Gelman et al. 2017). If we only use our understanding of the scale of the modeled parameters without any data-specific knowledge, we arrive at weakly-informative prior distributions. By weakly-informative I mean penalizing a-priori implausible values (e.g., a location parameter of 1000 on the logit-scale) without affecting the a-priori plausible parameter space too much (e.g., location parameters within the interval $[-3, 3]$ on the logit-scale). Weakly informative normal priors are often centered around $\mu = 0$ with σ appropriately chosen so that the prior covers the range of plausible parameter values but flattens out quickly outside of that space. For more details on priors for Bayesian IRT models, see the works of Bürkner (Bürkner 2019), Fox (Fox 2010), and Levy and Mislevy (Levy and Mislevy 2017).

A second class of priors for item parameters are hierarchical priors. For this purpose, we apply the non-centered parameterization of hierarchical models (Gelman et al. 2013) as detailed in the following. We split the linear predictor η (for any of the item parameter classes) into an overall parameter, $\bar{b}$, and an item-specific deviation from the overall parameter, $\tilde{b}_i$, so that

$$\eta_n = \bar{b} + \tilde{b}_{i_n} \tag{9}$$

Without additional constraints, this split is not identified as adding a constant to the overall parameter can be compensated by subtracting the same constant from all $\tilde{b}_i$ without changing the likelihood. In Bayesian multilevel models, we approach this problem by specifying a hierarchical prior on $\tilde{b}_i$ via

$$\tilde{b}_i \sim \text{normal}(0, \sigma) \tag{10}$$

where σ is the standard deviation parameter over items on the unconstrained scale. Importantly, not only $\tilde{b}_i$ but also the hyperparameters $\overline{b}$ and σ are estimated during the model fitting.

Using the prior distribution from (10), we would assume the item parameters of the same item to be unrelated but, in practice, it is quite plausible that they are intercorrelated (Bürkner 2019). To account for such (linear) dependency, we can extend Equation (10) to the multivariate case, so that we can model the vector $\tilde{b}_i = (\tilde{b}_{\beta_i}, \tilde{b}_{\alpha_i}, \tilde{b}_{\gamma_i}, \tilde{b}_{\psi_i})$ jointly via a multivariate normal distribution:

$$\tilde{b}_i \sim \mathrm{multinormal}(0, \boldsymbol{\sigma}, \Omega) \tag{11}$$

where $\boldsymbol{\sigma} = (\sigma_\beta, \sigma_\alpha, \sigma_\gamma, \sigma_\psi)$ is the vector of standard deviations and Ω is the correlation matrix of the item parameters (see also Bürkner 2017, 2019; Nalborczyk 2019). To complete the prior specification for the item parameters, we need to set priors on $\overline{b}$ and σ. For this purpose, weakly-informative normal prior on $\overline{b}$ and half-normal priors on σ are usually fine but other options are possible as well (see Bürkner 2019 for details).

A decision between hierarchical and non-hierarchical priors is not always easy. If in doubt, one can try out both kinds of priors and investigate whether they make a relevant difference. Personally, I prefer hierarchical priors as they imply some data-driven shrinkage due to their scale being learned by the model on the fly. In addition, they naturally allow item parameters to share information across parameter classes via the correlation matrix Ω.

With respect to the person parameters, it is most common to apply hierarchical priors of the form

$$\theta_j \sim \mathrm{normal}(0, \sigma_\theta) \tag{12}$$

where, similar as for hierarchical priors on item parameters, σ_θ is a standard deviation parameter estimated as part of the model on which we put a weakly-informative prior. To give the reader intuition: With the overall effects in our model, we model the probability that an average person (with an ability of zero, thus imagine the ability to be centered) answers an average item (with all item parameters at their average values which we estimate). The varying effects then give us the deviations from the average person or item, so that we can "customize" our prediction of the solution probability to more or less able persons, more or less easy items, more or less discriminatory items, etc.

In 2PL or more complex models, we can also fix σ_θ to some value (usually 1) as the scale is completely accounted for by the scale of the factor loadings σ_α. However, when using weakly-informative priors on both θ and α as well as on their hyperparameters, estimating σ_θ actually poses no problem for model estimation. Importantly, however, we do not include an overall person parameter $\overline{\theta}$ as done for item parameters in (9) as this would conflict with the overall location parameter $\overline{b}_\beta$ leading to substantial convergence problems in the absence very informative priors. This does not limit the model's usefulness as only differences of person parameters are of relevance, not their absolute values on an (in principal) arbitrary latent scale.

3. Analysis of the SPM-LS Data

The Bayesian IRT models presented above were applied to the SPM data of Myszkowski and Storme (Myszkowski and Storme 2018). The analyzed data consist of responses from 499 participants on the 12 most difficult SPM items and are freely available online (https://data.mendeley.com/datasets/h3yhs5gy3w/1). The data gathering procedure was described in detail by Myszkowski and Storme (Myszkowski and Storme 2018). Analyses were performed in R (R Core Team 2019) using brms (Bürkner 2017) and Stan (Carpenter et al. 2017) for model specification and estimation via MCMC. To investigate potential differences between hierarchical and non-hierarchical priors on the item parameters, models were estimated for both of these priors. Below, I refer to these approaches as hierarchical MCMC (MCMC-H) and non-hierarchical MCMC (MCMC-NH). Priors on person parameters were always hierarchical and weakly informative priors were imposed on the remaining parameters. All considered models converged well according to sample-agnostic (Vehtari 2019) and sampler-specific (Betancourt 2017) diagnostics. In the presentation of the results below, I omit details of prior distributions and auxiliary model fitting arguments. All details and the fully reproducible analysis are available on GitHub (https://github.com/paul-buerkner/SPM-IRT-models).

In addition to estimating the IRT models using MCMC, I also fitted the models via optimization as implemented in the mirt package (Chalmers 2012). Here, I considered two options: (1) a fully frequentist approach maximizing the likelihood under the same settings as in the original analysis of Myszkowski and Storme (Myszkowski and Storme 2018); and (2) a Bayesian optimization approach where I imposed the same priors on item parameters as in MCMC-NH. I refer to these two methods as maximum likelihood (ML) and maximum a-posteriori (MAP), respectively. For models involving latent variables, such as IRT models, ML or MAP optimization have to be combined with numerical integration over the latent variables as the mode of the joint distribution of all parameters including latent variables does not exist in general (e.g., see Bates et al. 2015). Such a combination of optimization and integration is commonly referred to as expectation-maximization (EM). A thorough discussion on EM methods is outside the scope of the present paper but the interested reader is referred to the work of Do and Batzoglou (Do and Batzoglou 2008).

3.1. Model Estimation

For estimation in a multilevel regression framework such as the one of brms, the data need to be represented in long format. In the SPM-LS data, the relevant variables are the binary response of the participants (variable `response2`) coded as either correct (`1`) or incorrect (`0`) as well as `person` and `item` identifiers. Following the principal of building models bottom-up, I start with the estimation of the most simple sensible model, that is, the 1PL model. When both person and item parameters are modeled hierarchically, the brms formula for the 1PL model can be specified as

```
formula_1pl <- bf(
formula = response2 ~ 1 + (1 | item) + (1 | person),
family = brmsfamily("bernoulli", link = "logit")
)
```

To apply non-hierarchical item parameters, we have to use the formula `response2 ~ 0 + item + (1 | person)` instead (see the code on Github for more details). For a thorough introduction and discussion of the brms formula syntax, see Bürkner (2017, 2018, 2019). As displayed in Figure 1, item parameter estimates of all methods are very similar for the 1PL model. In addition, their uncertainty estimates align closely as well. The brms formula for the 2PL model looks as follows:

```
formula_2pl <- bf(
  response2 ~ beta + exp(logalpha) * theta,
nl = TRUE,
  theta ~ 0 + (1 | person),
  beta ~ 1 + (1 |i| item),
  logalpha ~ 1 + (1 |i| item),
family = brmsfamily("bernoulli", link = "logit")
)
```

When comparing the formulas for the 1PL and 2PL models, we see that the structure has changed considerably as a result of going from a generalized linear model to a generalized non-linear model (see Bürkner 2019 for more details). As displayed in Figure 2, item parameter point and uncertainty estimates of all methods are rather similar for the 2PL model but not as close as for the 1PL model. In particular, we see that the slope estimates of Items 4 and 5 vary slightly, presumably due to different amounts of regularization implied by the priors. The brms formula for the 3PL model looks as follows:

```
formula_3pl <- bf(
  response2 ~ gamma + (1 - gamma) *
    inv_logit(beta + exp(logalpha) * theta),
nl = TRUE,
  theta ~ 0 + (1 | person),
  beta ~ 1 + (1 |i| item),
  logalpha ~ 1 + (1 |i| item),
  logitgamma ~ 1 + (1 |i| item),
nlf(gamma ~ inv_logit(logitgamma)),
family = brmsfamily("bernoulli", link = "identity"),
)
```

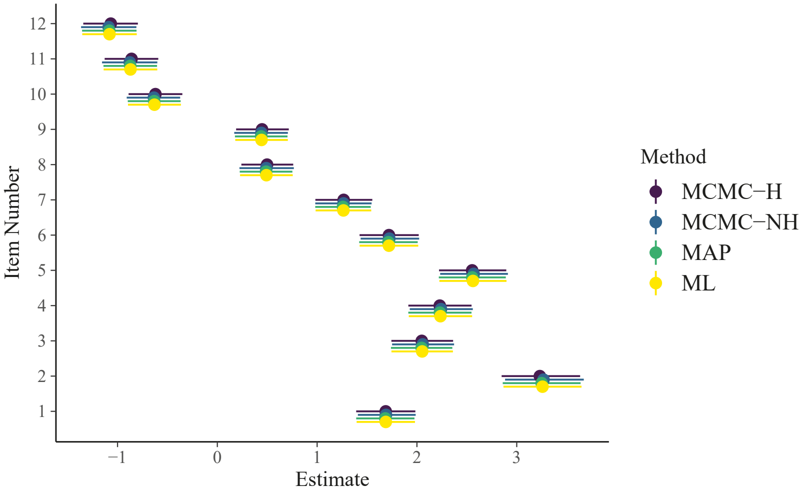


Figure 1. Item parameters of the 1PL model. Horizontal lines indicate 95% uncertainty intervals.

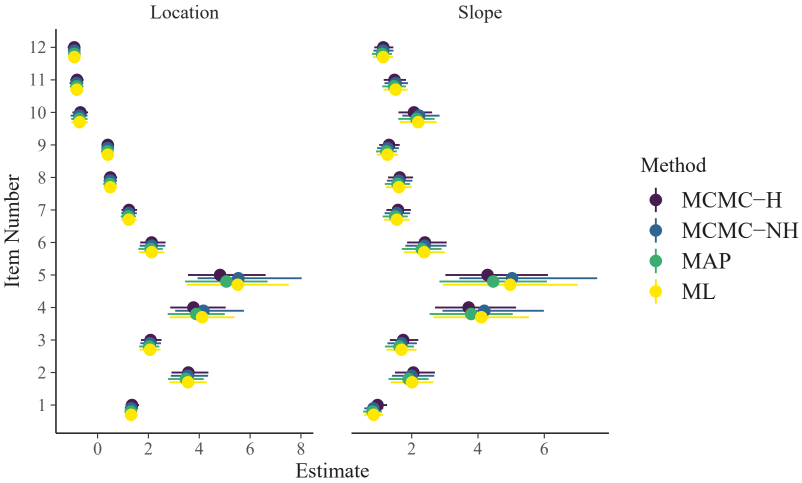


Figure 2. Item parameters of the 2PL model. Horizontal lines indicate 95% uncertainty intervals.

Note that, in the `family` argument, we now use `link = "identity"` instead of `link = "logit"` and build the logit link directly into the formula via `inv_logit(beta + exp(logalpha) * theta)`. This is necessary to correctly include guessing parameters (Bürkner 2019). As displayed in Figure 3, item parameter estimates of all methods are still quite similar when it comes to locations and slopes of the 3PL model. However, guessing parameter estimates are quite different: ML obtains point estimates of 0 for all but three items with uncertainty intervals ranging the whole definition space from 0 to 1. This is caused by an artifact in the computation of the approximate standard errors because point estimates are located at the boundary of the parameter space at which maximum likelihood theory does not hold. In contrast, point estimates of guessing parameters as obtained by all regularized models are close to but not exactly zero for most items and corresponding uncertainty estimates appear more realistic (i.e., much narrower) than those obtained by pure ML.

On Github, I also report results for the 3PL model with guessing probabilities fixed to 1/8 derived under the assumptions that, in the case of guessing, all alternatives are equally likely. According to Figure 3 and model comparisons shown on GitHub, this assumption does not seem to hold for the present data.

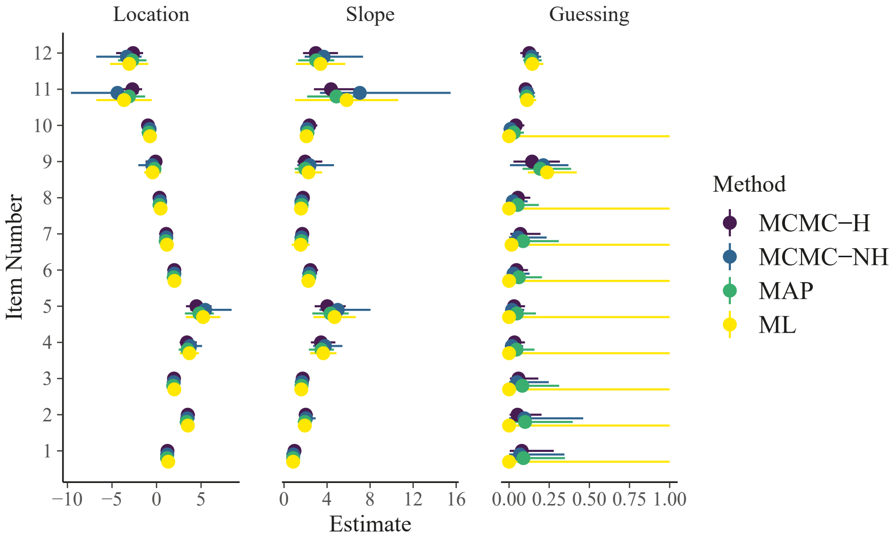


Figure 3. Item parameters of the 3PL model. Horizontal lines indicate 95% uncertainty intervals.

In Figure 4, I display person parameter estimates of the 3PL model. As we can see on the left-hand side of Figure 4, ML and MCMC-H point estimates align very closely. However, as displayed on the right-hand side of Figure 4, uncertainty estimates show some deviations, especially for more extreme point estimates (i.e., particularly good or bad performing participants). The brms formula for the 4PL model looks as follows:

```
formula_4pl <- bf(
  response2 ~ gamma + (1 - gamma - psi) *
    inv_logit(beta + exp(logalpha) * theta),
nl = TRUE,
  theta ~ 0 + (1 | person),
  beta ~ 1 + (1 |i| item),
  logalpha ~ 1 + (1 |i| item),
  logitgamma ~ 1 + (1 |i| item),
nlf(gamma ~ inv_logit(logitgamma)),
  logitpsi ~ 1 + (1 |i| item),
nlf(psi ~ inv_logit(logitpsi)),
family = brmsfamily("bernoulli", link = "identity")
)
```

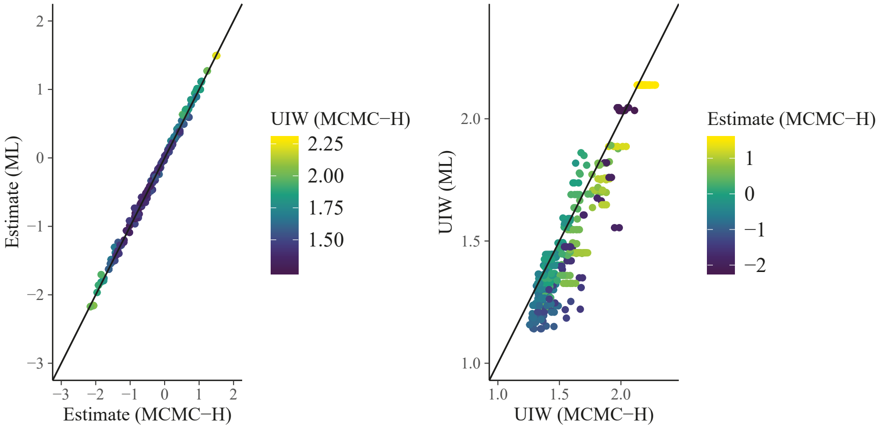


Figure 4. Comparison of 3PL person parameters: (Left) scatter plot of point estimates; and (Right) scatter plot of the associated 95% uncertainty interval widths (UIW).

As displayed in Figure 5, item parameter estimates of the 4PL model differ strongly from each other for different methods. In particular, ML point estimates were more extreme and no uncertainty estimates could be obtained due to singularity of the information matrix. It is plausible that the 4PL model is too difficult to be estimated based on the given data via ML without further regularization. Moreover, the estimates obtained by MCMC-H and MCMC-NH differ noticeably for some item parameters in the way that MCMC-NH estimates tend to be more extreme and uncertain as compared to MCMC-H. This suggests that, for these specifically chosen hierarchical and non-hierarchical priors, the former imply stronger regularization.

3.2. Model Comparison

Next, I investigate the required model complexity to reasonably describe the SPM data. For this purpose, I apply Bayesian approximate leave-one-out cross-validation (LOO-CV; ref. Vehtari et al. 2019; Vehtari et al 2017; Vehtari et al. 2018) as a method for model comparison, which is closely related to information criteria (Vehtari et al 2017). I only focus on the MCMC-H models here. Results for the MCMC-NH models are similar (see Github for details). As shown in Table 1, 3PL and 4PL models fit substantially better than the 1PL and 2PL models, while there was little difference between the former two. Accordingly, in the interest of parsimony, I would tend to prefer the 3PL model if a single model needed to be chosen. This coincides with the conclusions of Myszkowski and Storme (Myszkowski and Storme 2018).

Table 1. Bayesian Model comparison based on the leave-one-out cross-validation.

Model	ELPD	SE(ELPD)	ELPD-Difference	SE(ELPD-Difference)
4PL	−2544.7	42.6	0.0	0.0
3PL	−2547.8	42.8	−3.1	5.1
2PL	−2588.7	42.9	−44.0	9.5
1PL	−2655.0	43.8	−110.3	15.0

Note. ELPD, expected log posterior density; SE, standard error. Higher ELPD values indicate better model fit. ELPD differences are in comparison to the 4PL model.

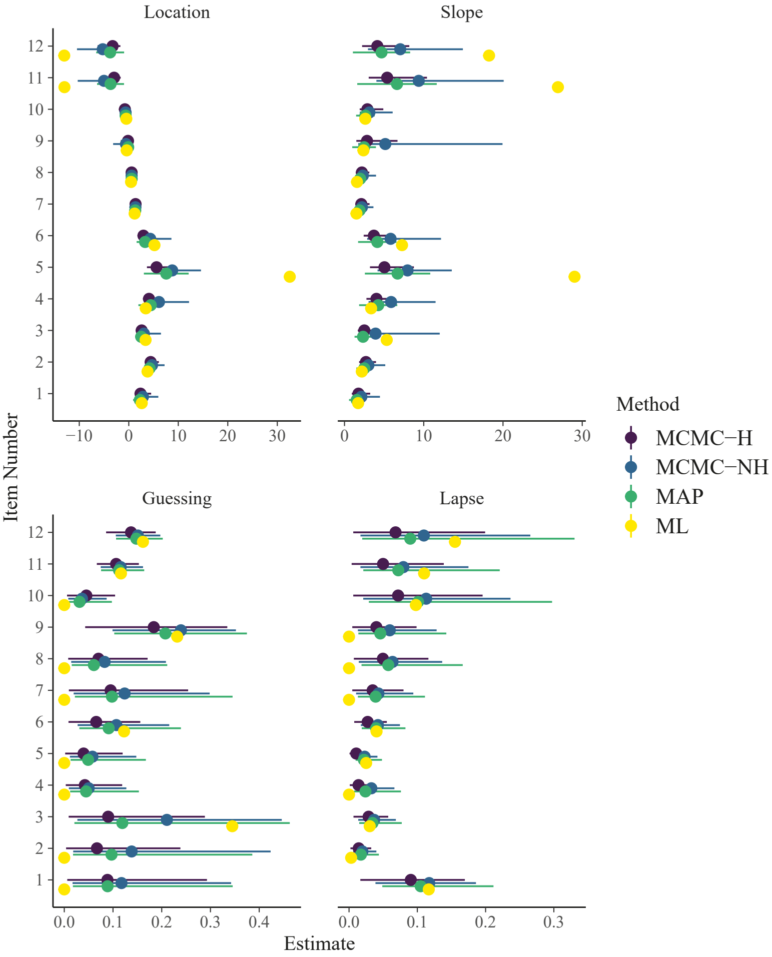


Figure 5. Item parameters of the 4PL model. Horizontal lines indicate 95% uncertainty intervals.

We can also investigate model fit using Bayesian versions of frequentist item or person fit statistics such as log-likelihood values (Glas and Meijer 2003). Independently of which statistic T is chosen, a Bayesian version of the statistic can be constructed as follows (Glas and Meijer 2003): First, the fit statistic is computed for the observed responses y. We denote it by $T(y,p)$, where $p = p(\theta, \xi)$ is the model implied response probability defined in Equation (3). As p depends on the model parameters, the posterior distribution over the parameters implies a posterior distribution over p, which in turn implies a posterior distribution over $T(y,p)$. Second, the fit statistic is computed for posterior predicted responses y_{rep} and we denote it by $T(y_{\text{rep}},p)$. Since y_{rep} reflects the (posterior distribution of) responses that would be predicted if the model was true, $T(y_{\text{rep}},p)$ provides a natural baseline for $T(y,p)$. Third, by comparing the posterior distributions of $T(y,p)$ and $T(y_{\text{rep}},p)$, we can detect item- or person-specific model misfit. In Figure 6, we show item-specific log-likelihood differences between predicted and observed responses for the 1PL model. It is clearly visible that the assumptions of the 1PL model are violated for almost half of the items. In contrast, the corresponding results for the 3PL model look much more reasonable (see Figure 7).

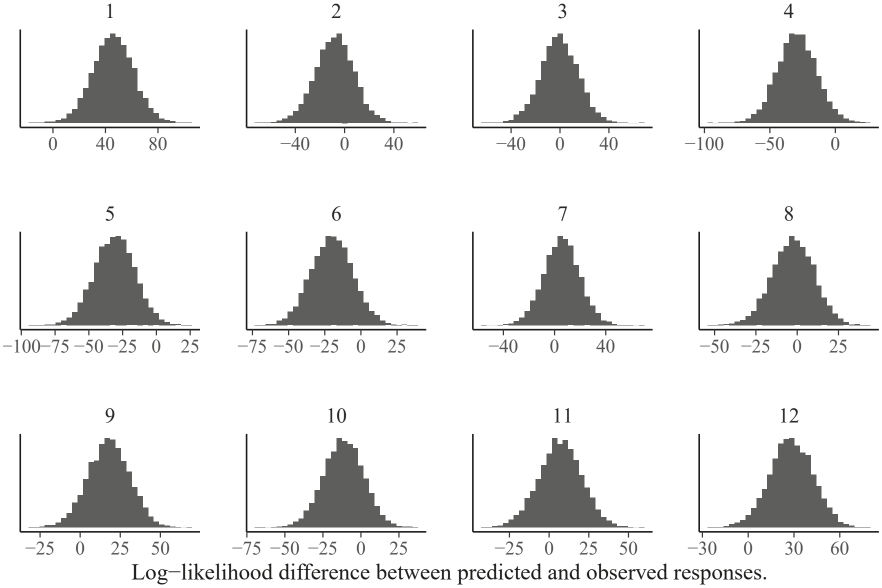


Figure 6. Item-specific posterior distributions of log-likelihood differences between predicted and observed responses for the 1PL model estimated via MCMC-H. If the majority of the posterior distribution is above zero, this indicates model misfit for the given item.

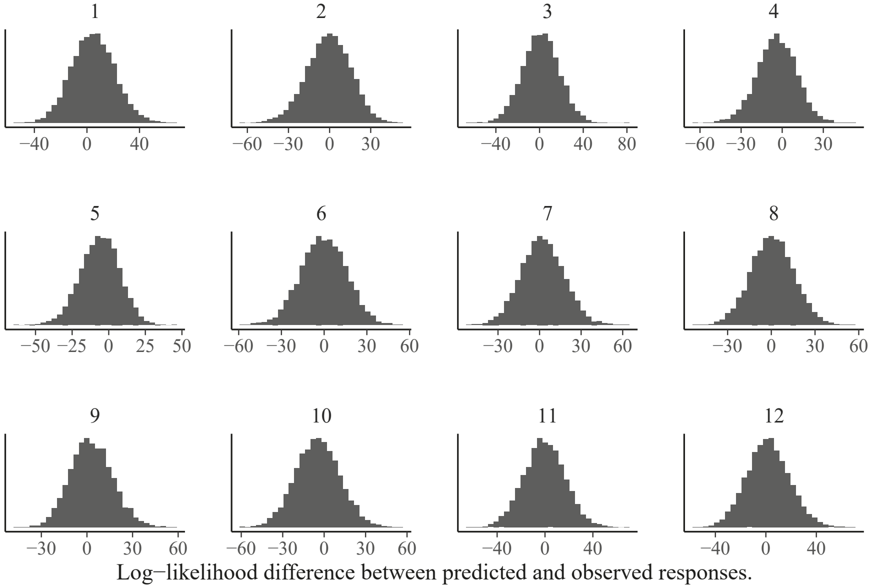


Figure 7. Item-specific posterior distributions of log-likelihood differences between predicted and observed responses for the 3PL model estimated via MCMC-H. If the majority of the posterior distribution is above zero, this indicates model misfit for the given item.

We can use the same logic to investigate person-specific model fit to find participants for whom the models do not make good predictions. In Figure 8, we show the predicted vs. observed log-likelihood differences of the 192nd person with response pattern $(0, 0, 1, 0, 0, 1, 0, 1, 1, 0, 0, 1)$. None of the models performs particularly well as this person did not answer some of the easiest items correctly (i.e., Items 2, 4, and 5) but was correct on some of the most difficult items (i.e., Items 8, 9, and 12). It is unclear what was driving such a response pattern. However, one could hypothesize that training effects over the course of the test played a role, which are not accounted for by all models presented here. To circumvent this in future test administrations, one could add more unevaluated practice items at the beginning of the test so that participants have the opportunity to become more familiar with the response format. Independently of the difference in model fit, person parameter estimates correlated quite strongly between different models and estimation approaches, with pairwise correlations exceeding $r = 0.97$ in all cases (see Figure 9 for an illustration).

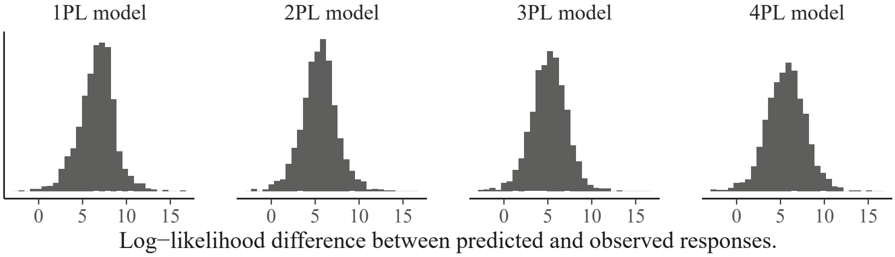


Figure 8. Person-specific posterior distributions of log-likelihood differences between predicted and observed responses for the 192nd person and different models estimated via MCMC-H. If the majority of the posterior distribution is above zero, this indicates model misfit for the given person.

The time required for estimation of the Bayesian models with brms via MCMC ranged from a couple of minutes for the 1PL model to roughly half an hour for the 4PL model (exact timings vary according to several factors, for instance, the number of iterations and chains, applied computing machines, or the amount of parallelization). In contrast, the corresponding optimization methods (ML and MAP) required only a few seconds for estimation in mirt. This speed difference of multiple orders of magnitude is typical for comparisons between MCMC and optimization methods (e.g., Bürkner 2017). Clearly, if speed is an issue for the given application, full Bayesian estimation methods via MCMC should be applied carefully.

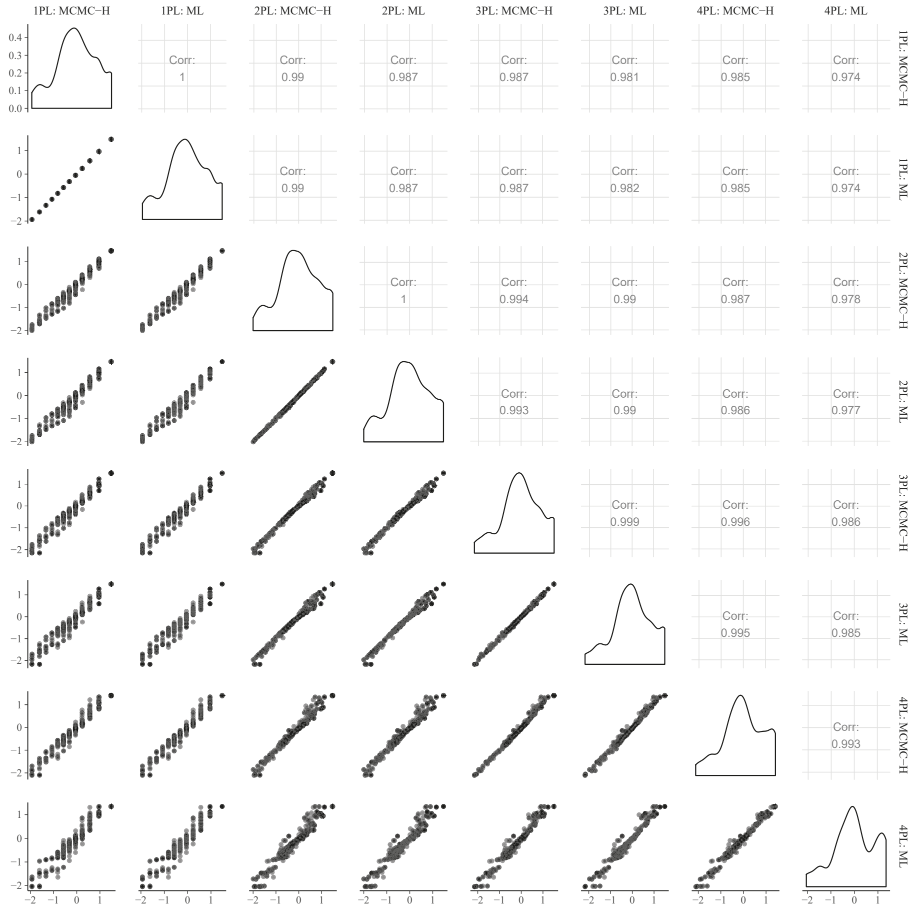


Figure 9. Scatter plots, bivariate correlations, and marginal densities of person parameters from MCMC-NH and ML models.

4. Discussion

In the present paper, I reanalyze data to validate a short version of Standard Progressive Matrices (SPM-LS; Myszkowski and Storme 2018) using Bayesian IRT models. By comparing out-of-sample predictive performance, I found evidence that the 3PL model with estimated guessing parameters outperformed simpler models and performed similarly well as the 4PL model, which additionally estimated lapse parameters. As specifying and fitting the 4PL model is substantially more involved than the 3PL model without apparent gains in out-of-sample predictive performance, I argue that the 3PL model should probably be the model of choice within the scope of all models considered here. That is, I come to a similar conclusion as Myszkowski and Storme (Myszkowski and Storme 2018) in their original analysis despite using different frameworks for model specification and estimation (Bayesian vs. frequentist) as well as predictive performance (approximate leave-one-out cross-validation (Vehtari et al 2017) vs. corrected AIC and χ^2-based measures (Maydeu-Olivares 2013).

Even though I reach the same conclusions as Myszkowski and Storme (Myszkowski and Storme 2018) reached with conventional frequentist methods, I would still like to point out some advantages of applying Bayesian methods that we have seen in this application. With regard to item parameters, Bayesian and frequentist estimates showed several important differences for the most complex 3PL and 4PL IRT models. First, point estimates of items with particularly high difficulty or slope were more extreme in the frequentist maximum likelihood estimation. One central reason is the use weakly informative priors in the Bayesian models which effectively shrunk extremes a little towards the mean thus providing more conservative and robust estimates (Gelman et al. 2013). Specifically, for the 4PL model, the model structure was also too complex to allow for reasonable maximum likelihood estimation in the absence of any additional regularization to stabilize inference. The latter point also becomes apparent because no standard errors of the ML estimated items parameters in the 4PL model could be computed due to singularity of the information matrix. Even when formally computable, uncertainty estimates provided by the frequentist IRT models were not always meaningful. For instance, in the 3PL model, the confidence intervals of guessing parameters estimated to be close to zero were ranging the whole definition space between zero and one. This is clearly an artifact as maximum likelihood theory does not apply at the boundary of the parameter space and hence computation of standard errors is likely to fail. As such, these uncertainty estimates should not be trusted. Robust alternatives to computing approximate standard errors via maximum likelihood theory are bootstrapping or other general purpose data resampling methods (e.g., Freedman 1981; Junker and Sijtsma 2001; Mooney et al. 1993). These resampling methods come with additional computational costs as the model has to be repeatedly fitted to different datasets but can be used even in problematic cases where standard uncertainty estimators fail.

In contrast, due to the use of weakly informative priors, the Bayesian models provided sensible uncertainty estimates for all item parameters of every considered IRT model. MCMC and MAP estimates provided quite similar results for the item parameters in the context of the SPM-LS data and applied binary IRT models. However, there is no guarantee that this will be generally the case and thus it is usually safer to apply MCMC methods when computationally feasible. In addition, for the propagation of uncertainty to new quantities, for instance, posterior predictions, MCMC or other sampling-based methods are required. In the case study, I demonstrated this feature in the context of item and person fit statistics, which revealed important insides into the assumptions of the applied IRT models.

With regard to point estimates of person parameters, I found little differences between all considered Bayesian and frequentist IRT models. Pairwise correlations between point estimates of two different models were all exceeding $r = 0.97$ and often even larger than $r = 0.99$. This should not imply, however, that the model choice does not matter in the context of person parameter estimation (Loken and Rulison 2010). Although point estimates were highly similar, uncertainty estimates of person parameters varied substantially across model classes. Thus, it is still important to choose an appropriately complex model for the data (i.e., the 3PL model in our case) in order to get sensible uncertainty estimates. The latter are not only relevant for individual diagnostic purposes, which is undoubtedly a major application of intelligence tests, but also when using person parameters as predictors in other models while taking their estimation uncertainty into account. In addition, uncertainty estimates of Bayesian and frequentist models varied substantially even within the same model class, in particular for 3PL and 4PL models. Without a known ground truth, we have no direct evidence which of the uncertainty estimates are more accurate (with respect to some Bayesian and/or frequentist criteria), but I would argue in favor of the Bayesian results as they should have benefited from the application of weakly informative priors and overall more robust inference procedures for the considered classes of models. Overall, it is unsurprising that Bayesian methods have an easier time estimating uncertainty as it is more natural to do so in a Bayesian framework. We have also seen the important advantage of Bayesian methods that is their ability to more easily accommodate more complex models. However, we have also seen that, for simpler models, Bayesian and frequentist methods provide very similar results, which really speaks in favor of both methods and should highlight for the reader that both choices are valid options in this case and neither should be attacked. Developing this understanding seems necessary with the increased application of Bayesian methods and the accompanying arguments of whether this is a valid option.

The analysis presented here could be extended in various directions. First, one could fit polytomous IRT models that take into account potential differences between distractors and thus use more information than binary IRT models. Such polytomous IRT models were also fitted by Myszkowski and Storme (Myszkowski and Storme 2018) and demonstrated some information gain as compared to their binary counterparts. Fitting these polytomous IRT models in a Bayesian framework is possible as well, but currently not supported by brms in the here required form. Instead, one would have to use Stan directly, or another probabilistic programming language, whose introduction is out of scope of the present paper. Second, one could consider multiple person traits/latent variables to investigate the unidimensionality of the SPM-LS test. Currently, this cannot be done in brms in an elegant manner but will be possible in the future once formal measurement models have been implemented. For the time being, one has to fall back to full probabilistic programming languages such as Stan or more specialized IRT software that supports multidimensional Bayesian IRT models. According to Myszkowski and Storme (Myszkowski and Storme 2018), the SPM-LS test is sufficiently unidimensional to justify the application of unidimensional IRT models. Accordingly, the lack of multidimensional models does not constitute a major limitation for the present analysis.

In summary, I was able to replicate several key findings of Myszkowski and Storme (Myszkowski and Storme 2018). Additionally, I demonstrated that Bayesian IRT models have some important advantages over their frequentist counterparts when it comes to reliably fitting more complex response processes and providing sensible uncertainty estimates for all model parameters and other quantities of interest.

Funding: This research received no external funding.

Acknowledgments: I want to thank Marie Beisemann and two anonymous reviewers for their valuable comments on earlier versions of the paper. To conduct the presented analyses and create this paper, I used the programming language R (R Core Team 2019) through the interface RStudio (R Studio Team 2018). Further, the following R packages were crucial (in alphabetical order): brms (Bürkner 2017), dplyr (Wickham et al. 2019), GGally (Schloerke et al. 2018), ggplot2 (Wickham 2016), kableExtra (Zhu 2019), knitr (Xie 2015), loo (Vehtari et al 2017), mirt (Chalmers 2012), papaja (Aust 2018), patchwork (Pedersen 2017), rmarkdown (Xie et al. 2018), rstan (Carpenter et al. 2017), and tidyr (Wickham and Henry 2019).

Conflicts of Interest: The authors declare no conflicts of interest.

References

Ackerman, Phillip L., and Ruth Kanfer. 2009. Test length and cognitive fatigue: An empirical examination of effects on performance and test-taker reactions. *Journal of Experimental Psychology: Applied* 15: 163. [CrossRef]

Allison, J. Ames, and Chi Hang Au. 2018. Using Stan for item response theory models. *Measurement: Interdisciplinary Research and Perspectives* 16: 129–34.

Aust, F., and M. Barth. 2018. Papaja: Create APA Manuscripts with R Markdown. Available online: https://github.com/crsh/papaja (accessed on 3 February 2020).

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Soft-Ware* 67: 1–48. [CrossRef]

Betancourt, Michael. 2017. A conceptual introduction to Hamiltonian Monte Carlo. *arXiv* arXiv:1701.02434.

Bürkner, Paul-Christian. 2017. brms: An R package for bayesian multilevel models using Stan. *Journal of Statistical Software* 80: 1–28. [CrossRef]

Bürkner, Paul-Christian. 2018. Advanced Bayesian multilevel modeling with the R package brms. *The R Journal* 10: 395–411. [CrossRef]

Bürkner, Paul-Christian. 2019. Bayesian item response modelling in R with brms and Stan. *arXiv* arXiv:1905.09501.

Carpenter, Bob, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2017. Stan: A probabilistic programming language. *Journal of Statistical Software* 76: 1–32. [CrossRef]

Chalmers, R. Philip. 2012. mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software* 48: 1–29. [CrossRef]

Culpepper, Steven Andrew. 2016. Revisiting the 4-parameter item response model: Bayesian estimation and application. *Psychometrika* 81: 1142–63. [CrossRef]

Culpepper, Steven Andrew. 2017. The prevalence and implications of slipping on low-stakes, large-scale assessments. *Journal of Educational and Behavioral Statistics* 42: 706–25. [CrossRef]

Curtis, S. McKay. 2010. BUGS code for item response theory. *Journal of Statistical Software* 36: 1–34. [CrossRef]

Depaoli, Sarah, James P. Clifton, and Patrice R. Cobb. 2016. Just another Gibbs sampler (JAGS) flexible software for MCMC implementation. *Journal of Educational and Behavioral Statistics* 41: 628–49. [CrossRef]

Do, Chuong B., and Serafim Batzoglou. 2008. What is the expectation maximization algorithm? *Nature Biotechnology* 26: 897. [CrossRef] [PubMed]

Embretson, Susan E., and Steven P. Reise. 2013. *Item Response Theory*. Hove: Psychology Press.

Fox, Jean-Paul. 2010. *Bayesian Item Response Modeling: Theory and Applications*. Berlin/Heidelberg: Springer.

Freedman, David A. 1981. Bootstrapping regression models. *The Annals of Statistics* 9: 1218–28. [CrossRef]

Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. 2013. *Bayesian Data Analysis*, 3rd ed. Boca Raton: Chapman Hall/CRC. [CrossRef]

Gelman, Andrew, Daniel Simpson, and Michael Betancourt. 2017. The prior can often only be understood in the context of the likelihood. *Entropy* 19: 555–67. [CrossRef]

Glas, Cees A. W., and Rob R. Meijer. 2003. A Bayesian approach to person fit analysis in item response theory models. *Applied Psychological Measurement* 27: 217–33. [CrossRef]

Jensen, Arthur R., Dennis P. Saccuzzo, and Gerald E. Larson. 1988. Equating the standard and advanced forms of the Raven progressive matrices. *Educational and Psychological Measurement* 48: 1091–95. [CrossRef]

Junker, Brian W., and Klaas Sijtsma. 2001. Nonparametric item response theory in action: An overview of the special issue. *Applied Psychological Measurement* 25: 211–20. [CrossRef]

Levy, Roy, and Robert J. Mislevy. 2017. *Bayesian Psychometric Modeling*. Boca Raton: Chapman Hall/CRC.

Loken, Eric, and Kelly L. Rulison. 2010. Estimation of a four-parameter item response theory model. *British Journal of Mathematical and Statistical Psychology* 63: 509–25. [CrossRef]

Lord, Frederic M. 2012. *Applications of Item Response Theory to Practical Testing Problems*. Milton Park: Routledge.

Lunn, David, David Spiegelhalter, Andrew Thomas, and Nicky Best. 2009. The BUGS project: Evolution, critique and future directions. *Statistics in Medicine* 28: 3049–67. [CrossRef] [PubMed]

Luo, Yong, and Hong Jiao. 2018. Using the Stan program for bayesian item response theory. *Educational and Psychological Measurement* 78: 384–408. [CrossRef] [PubMed]

Maydeu-Olivares, Alberto. 2013. Goodness-of-fit assessment of item response theory models. *Measurement: Interdisciplinary Research and Perspectives* 11: 71–101. [CrossRef]

Mooney, Christopher F., Christopher L. Mooney, Christopher Z. Mooney, Robert D. Duval, and Robert Duvall. 1993. *Bootstrapping: A Nonparametric Approach to Statistical Inference*. Thousand Oaks: Sage.

Myszkowski, Nils, and Martin Storme. 2018. A snapshot of g? Binary and polytomous item-response theory investigations of the last series of the standard progressive matrices (SPM-LS). *Intelligence* 68: 109–16. [CrossRef]

Nalborczyk, Ladislas, Cédric Batailler, Hélène Lœvenbruck, Anne Vilain, and Paul-Christian Bürkner. 2019. An introduction to bayesian multilevel models using brms: A case study of gender effects on vowel variability in standard indonesian. *Journal of Speech, Language, and Hearing Research* 62: 1225–42. [CrossRef]

Pedersen, Thomas Lin. 2017. Patchwork: The Composer of Ggplots. Available online: https://github.com/thomasp85/patchwork (accessed on 3 February 2020).

Pind, Jörgen, Eyrún K. Gunnarsdóttir, and Hinrik S. Jóhannesson. 2003. Raven's standard progressive matrices: New school age norms and a study of the test's validity. *Personality and Individual Differences* 34: 375–86. [CrossRef]

Plummer, Martyn. 2013. JAGS: Just Another Gibbs Sampler. Available online: http://mcmc-jags.sourceforge.net/ (accessed on 3 February 2020).

Rasch, Georg. 1961. On general laws and the meaning of measurement in psychology. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley: University of California Press, vol. 4, pp. 321–33.

Raven, John C. 1941. Standardization of progressive matrices, 1938. *British Journal of Medical Psychology* 19: 137–50. [CrossRef]

R Core Team. 2019. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available online: https://www.R-project.org/ (accessed on 3 February 2020).

Robitzsch, Alexander. 2019. Sirt: Supplementary Item Response Theory Models. Available online: https://CRAN.R-project.org/package=sirt (accessed on 3 February 2020).

Robitzsch, Alexander, Thomas Kiefer, and Margaret Wu. 2019. TAM: Test Analysis Modules. Available online: https://CRAN.R-project.org/package=TAM (accessed on 3 February 2020).

R Studio Team. 2018. *RStudio: Integrated Development for R*. Boston: RStudio, Inc., vol. 42.

Rupp, Andre A., Dipak K. Dey, and Bruno D. Zumbo. 2004. To Bayes or not to Bayes, from whether to when: Applications of Bayesian methodology to modeling. *Structural Equation Modeling* 11: 424–51. [CrossRef]

Schloerke, Barret, Jason Crowley, Di Cook, Heike Hofmann, Hadley Wickham, François Briatte, Moritz Marbach, Edwin Thoen, Amos Elberg, and Joseph Larmarange. 2018. GGally: Extension to 'ggplot2'. Available online: https://CRAN.R-project.org/package=GGally (accessed on 3 February 2020).

van der Linden, Wim J., and Ronald K. Hambleton, eds. 1997. *Handbook of Modern Item Response Theory*. Berlin/Heidelberg: Springer.

Vehtari, Aki, Daniel Simpson, Andrew Gelman, Yuling Yao, and Jonah Gabry. 2019. Pareto smoothed importance sampling. *arXiv* arxiv:1507.02646.

Vehtari, Aki, Andrew Gelman, and Jonah Gabry. 2017. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing* 27: 1413–32. [CrossRef]

Vehtari, Aki, Andrew Gelman, and Jonah Gabry. 2018. Loo: Efficient Leave-One-Out Cross-Validation and WAIC for Bayesian Models. Available online: https://github.com/stan-dev/loo (accessed on 3 February 2020).

Vehtari, Aki, Andrew Gelman, Daniel Simpson, Bob Carpenter, and Paul-Christian Bürkner. 2019. Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of MCMC. *arXiv* arXiv:1903.08008.

Waller, Niels G., and Leah Feuerstahler. 2017. Bayesian modal estimation of the four-parameter item response model in real, realistic, and idealized datasets. *Multivariate Behavioral Research* 52: 350–70. [CrossRef] [PubMed]

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. New York: Springer. Available online: http://ggplot2.org (accessed on 3 February 2020).

Wickham, Hadley, Romain Francois, Lionel Henry, and Kirill Müller. 2019. Dplyr: A Grammar of Data Manipulation. Available online: https://CRAN.R-project.org/package=dplyr (accessed on 3 February 2020).

Wickham, H., and L. Henry. 2019. Tidyr: Tidy Messy Data. Available online: https://CRAN.R-project.org/package=tidyr (accessed on 3 February 2020).

Xie, Yihui. 2015. *Dynamic Documents with R and Knitr*, 2nd ed. Boca Raton: Chapman Hall/CRC. Available online: https://yihui.name/knitr/ (accessed on 3 February 2020).

Xie, Yihui, Joseph J. Allaire, and Garrett Grolemund. 2018. *R Markdown: The Definitive Guide*. Boca Raton: Chapman Hall/CRC. Available online: https://bookdown.org/yihui/rmarkdown (accessed on 3 February 2020).

Zhan, Peida, Hong Jiao, Kaiwen Man, and Lijun Wang. 2019. Using JAGS for Bayesian cognitive diagnosis modeling: A tutorial. *Journal of Educational and Behavioral Statistics*. [CrossRef]

Zhu, Hao. 2019. kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax. Available online: https://CRAN.R-project.org/package=kableExtra (accessed on 3 February 2020).

Journal of
Intelligence

MDPI

Article

How Much *g* Is in the Distractor? Re-Thinking Item-Analysis of Multiple-Choice Items

Boris Forthmann *, Natalie Förster, Birgit Schütze, Karin Hebbecker, Janis Flessner, Martin T. Peters and Elmar Souvignier

Institute of Psychology in Education, University of Münster, 48149 Münster, Germany;
natalie.foerster@uni-muenster.de (N.F.); birgit.schuetze@uni-muenster.de (B.S.);
karin.hebbecker@uni-muenster.de (K.H.); flessner@uni-muenster.de (J.F.);
martin.peters@uni-muenster.de (M.T.P.); elmar.souvignier@uni-muenster.de (E.S.)
* Correspondence: boris.forthmann@wwu.de; Tel.: +49-251-83-34162

Received: 31 January 2020; Accepted: 3 March 2020; Published: 9 March 2020

check for updates

Abstract: Distractors might display discriminatory power with respect to the construct of interest (e.g., intelligence), which was shown in recent applications of nested logit models to the short-form of Raven's progressive matrices and other reasoning tests. In this vein, a simulation study was carried out to examine two effect size measures (i.e., a variant of Cohen's ω and the canonical correlation R_{CC}) for their potential to detect distractors with ability-related discriminatory power. The simulation design was adopted to item selection scenarios relying on rather small sample sizes (e.g., $N = 100$ or $N = 200$). Both suggested effect size measures (Cohen's ω only when based on two ability groups) yielded acceptable to conservative type-I-error rates, whereas, the canonical correlation outperformed Cohen's ω in terms of empirical power. The simulation results further suggest that an effect size threshold of 0.30 is more appropriate as compared to more lenient (0.10) or stricter thresholds (0.50). The suggested item-analysis procedure is illustrated with an analysis of twelve Raven's progressive matrices items in a sample of $N = 499$ participants. Finally, strategies for item selection for cognitive ability tests with the goal of scaling by means of nested logit models are discussed.

Keywords: Raven's progressive matrices; intelligence; distractors; item analysis

1. Introduction

Distractors are a fundamental part of the item content in multiple-choice items (Thissen et al. 1989; Guttman and Schlesinger 1967). That fact is taken into account in both traditional and contemporary distractor analysis (Gierl et al. 2017). An approach that falls in the category of contemporary distractor analysis is Myszkowski and Storme (2018) nested logit model application to the latest short form of Raven's Progressive Matrices. The nested logit model family (Suh and Bolt 2010) concurrently uses accuracy and distractor choice information from each item to improve ability estimation. That is, item responses to multiple-choice items are modeled in terms of solution behavior (i.e., solved vs. not-solved) by means of a logistic item response theory (IRT) model for binary items (e.g., 1PL, 2PL or 3PL) at the first place. Then, given the item has not been solved, distractor choices are modeled by Bock's nominal response model (NRM) (Bock 1972). Hence, nested logit models, as used by Myszkowski and Storme (2018), include varying discrimination parameters for each distractor. Traditional distractor analysis, as part of a thorough item analysis, does not necessarily focus on this aspect of distractor choices.

The primary focus on solution behavior and a secondary focus on distractor choices is one advantage of nested logit models for the modeling of figural matrix items as compared to other polytomous IRT models. Indeed, there is clear evidence in the literature that using constructive matching [i.e., a strategy focused on constructing the correct solution (Snow 1980; Bethell-Fox et al. 1984)] is positively correlated with cognitive ability. This was found for constructive matching indicators (e.g., self-reported strategy use, estimated latent strategy classes, or the proportion of overall time spent on the item content) derived from the paperfolding test (Snow 1980), figural analogies (Bethell-Fox et al. 1984; Schiano et al. 1989), and figural matrices (Vigneau et al. 2006; Mitchum and Kelley 2010; Hayes et al. 2011; Gonthier and Thomassin 2015; Gonthier and Roulin 2019). In addition, analogous indicators for usage of distractor elimination strategies (e.g., the proportion of overall time spent on the response alternatives or back and forth eye movements between the item content and the response alternatives) were found to be negatively correlated with test performance (Bethell-Fox et al. 1984; Schiano et al. 1989; Vigneau et al. 2006; Hayes et al. 2011; Jarosz and Wiley 2012; Arendasy and Sommer 2013; Gonthier and Thomassin 2015; Gonthier and Roulin 2019). In line with these findings, (Myszkowski and Storme 2018) pointed out that nested logit models which take into account solution behavior, as well as distractor choice, are perhaps best suited to model solution processes starting with constructive matching, and given that a solution, cannot be reached, shifting towards distractor elimination strategies at a later stage. Indeed, Gonthier and Roulin (2019) reported results in line with the idea that both constructive matching and response elimination might be used on the same item.

In addition, the idea that distractor choice provides useful psychometric information existed even before the invention of IRT models for polytomous scoring (Guttman and Schlesinger 1967; Davis and Fifer 1959), and hence, informativeness of distractors has been studied by other approaches than IRT. This is evident in early studies that examined gender in relation to certain error patterns, such as failing to discriminate between the correct option and a distractor designed by rotating the correct solution (Sigel 1963; Vejleskov 1968). In a similar vein, Jacobs and Vandeventer (1970) found that the proportion of choosing distractors that either take into account solely the horizontal or solely the vertical facet was positively correlated with performance on the Coloured Progressive Matrices. Moreover, Vodegel Matzen et al. (1994) found that choosing distractors that share features with the solution or distractors that were a repetition of one of the adjacent entries to the missing element in the matrix discriminated best between children with varying levels of performance [for a complete overview of studies focusing on error analysis in figural matrix items see Kunda et al. (2016)]. Finally, IRT approaches were also found to reveal discriminatory power of distractors with respect to ability (Myszkowski and Storme 2018; Thissen 1976; Storme et al. 2019).

To sum up, evidence on strategy use and informativeness of distractors in figural matrix items seemingly adhere to the idea behind nested logit models. Hence, in combination with the use of rule-based distractor generation (Guttman and Schlesinger 1967; Hornke and Habon 1986; Matzen et al. 2010; Blum et al. 2016; Blum and Holling 2018) to construct items with discriminating distractors, this item family appears to be promising for test development based on nested logit models. In particular, this allows the construction of tests with higher measurement precision at the lower end of the ability range because differentiated information about the ability of those who did not solve the item is taken from distractor choices (Myszkowski and Storme 2018; Storme et al. 2019). To date, however, proper distractor evaluation tools for such item development are not available. In addition, direct use of nested logit models with small sample sizes (e.g., $N = 200$) seems not feasible in light of the many parameters that need to be estimated. In this vein, it is especially unclear which descriptive statistics are informative to allow item pre-selection based on criteria in line with the idea of distractor discrimination at an early stage in the item selection process when candidate items are tested in small samples. Thus, the goal of the current work is to examine Cohen's ω (Cohen 1992) based on ability groups and distractor choice and the canonical correlation (Thompson 1984; Klecka 1980) between test performance and distractor choice for their potential to detect items with discriminatory distractors. First, we review the potential of traditional distractor analysis tools to reveal the discriminatory power of an item's distractor set and propose Cohen's ω and the canonical correlation as useful effect sizes that correspond with similar approaches found in the literature. Second, we evaluate how well these effect sizes perform as a detection method for item pre-selection in a simulation study. Finally, we apply the proposed effect sizes for distractor analysis to the dataset around which this special issue is organized to evaluate the distractors' potential to be used for nested logit models.

1.1. Distractor Analysis as Part of Traditional Analysis of Multiple-Choice Items

Item analysis refers to a set of descriptive statistics that are useful during the process of developing an item pool for a new psychological test (e.g., for the measurement of intelligence). These statistics are most often used in pilot studies in which the item pool (or a subset thereof) is administered to a relatively small sample (say $N = 50$ to $N = 300$) for the purpose of informed item selection. In this process, item analysis is used to provide the first evidence of each item's psychometric properties, such as difficulty, dimensionality, or discrimination (Henrysson 1971). In this work, we will focus on item discrimination which refers to the relationship between a person's ability and item performance (Lord 1980; Yen and Fitzpatrick 2006). A highly discriminating item results in a higher solving probability for persons with higher ability level and in lower solving probability for persons with lower ability level (i.e., as compared to a low discriminating item in which solving probabilities are more evenly distributed across the range of ability levels). More specifically, we will focus on ability-related discriminatory power of distractors in multiple-choice items, but not at the level of solution behavior (i.e., accuracy). That is, in case that an item has not been solved correctly by a test-taker it might be the case that choosing a particular distractor vs. choosing one of the other distractors is more likely for persons with higher ability, whereas, persons with lower ability are less likely to choose this distractor. Distractor discrimination parameters are indeed included in certain IRT models, such as the NRM or nested logit models. However, here we focus on more simple item effect sizes that can potentially reveal if items have discriminative distractors in pilot studies which usually have small sample sizes. Pilot studies for scale development can have goals of varying complexity. For example, the smallest sample sizes have been proposed for very early checks of instruction or item wordings (Johanson and Brooks 2010). Pilot studies with a focus on more complex properties of psychological tests, such as latent variable profiles (Von der Embse et al. 2014), for example, may have even sample sizes as large as 1000 participants. For the purpose of this work, we consider reasonably large sample sizes that reflect practice in cognitive ability research (Arendasy et al. 2006). On this basis, items can be selected for a test that, in the next step, could be scaled by means of a nested logit model. This would increase the reliability for low ability test takers (Myszkowski and Storme 2018; Suh and Bolt 2010; Storme et al. 2019).

Perhaps the best-known index of discrimination in item analysis is the item-scale correlation (Henrysson 1971; Cureton 1966) and its various corrections for part-whole overlap (Henrysson 1971) or lack of reliability (Cureton 1966). Conceptually, the same information is captured by factor loadings in factor analytical methods (Henrysson 1962) and discrimination parameters included in IRT models, such as the 2PL to 4PL models (Barton and Lord 1981) or the generalized partial credit model (Muraki 1992), for example. Hence, simple item-scale correlations can be considered as the pilot study counterpart of model parameters included in more complex approaches used for the final scaling of the test. For traditional item analysis, cut-offs exist to decide if items from a pilot study are retained in the item pool for final test calibration. Several suggestions for such cut-offs have been made in the literature, such as item-scale correlations >0.30 (Nunnally and Bernstein 1994) or at least >0.20 (Crocker and Algina 1986). In addition, similar cut-offs exist for standardized factor loadings (Kline 2000) taken from factor analytical approaches which are also used in pilot studies. However, comparable complementary sets of item statistics and model parameters for item selection and final scaling of the test, respectively, along with commonly used item-scale correlation cut-offs for item selection are not available for items with potentially discriminative distractors.

To illustrate traditional distractor analysis statistics, we created two example datasets, and used the first item from each of these datasets. To facilitate illustration below, we simply refer to the first item from the first dataset to as Example-Item 1 and to the first item from the second dataset to as Example-Item 2. These two items were simulated with three distractors and according to the same population model despite the distractor discrimination parameters. For both items, correct solution behavior was modeled by means of the 2PL with moderate discrimination and difficulty parameters ranging from −1.15 to −0.85 for a set of *difficult* items (see the setup for the simulation study in Section 2.1.2). When participants did not solve the item, their distractor choice was simulated according to NRM intercept parameters as described below, and for Example-item 1 the discrimination parameters were fixed at zero, whereas, for Example-Item 2, high discrimination parameters were simulated. Software code to replicate these data are available in the Open Science Framework repository for this work (https://osf.io/9tp8h/). All statistics introduced below are illustrated for these two items in Table 1 and interpreted in more detail in the respective sections and subsections below.

Table 1. Distractor specific statistics for Example-item 1 and Example-item 2.

Distractor	**Relative Choice Frequency < 0.05**	PB_D	PB_{DC}	ω_D	γ
	Item 1/Item 2	Item 1/Item 2	Item 1/Item 2	Item 1/Item 2	Item 1/Item 2
Distractor 1	0/0	−0.18/−0.21	−0.54/−0.58	0.57/0.68	1/1
Distractor 2	0/0	−0.29/−0.08	−0.56/−0.47	0.55/0.39	1/1
Distractor 3	0/0	−0.13/−0.36	−0.50/−0.68	0.58/0.97	1/1

Relative choice frequency < 0.05 = Number of distractors with a relative choice frequency below 0.05; PB_D = point-biserial correlation for the contrast between participants who chose D vs. participants who chose any other option (including the correct option) with respect to test performance; PB_{DC} = point-biserial correlation for the contrast between participants who chose D vs. participants who chose the correct option with respect to test performance; ω_D—Haladyna-Downing approach (Haladyna and Downing 1993) = Cohen's ω based on choice frequencies restricted to D as a function of 5 ability groups based on equi-distant quantiles; γ = Goodman-Kruskal γ for the relationship between test performance based on all other items and the probability to choose the correct response as estimated based on a $2 \times J$ contingency table (with J is the number of possible performance scores) which has been suggested by (Love 1997) as an index for the evaluation of rising selection ratios.

1.1.1. Distractor Choice Frequency

The distractor choice frequency is often applied as the first criterion for distractor evaluation with the recommendation that useful distractors should be chosen by at least 5% of the participants (Haladyna and Downing 1993). Distractors not fulfilling this criterion in a pilot study would be considered as non-functioning and need revision. Hence, we conclude that distractors in pilot studies should in the first place pass the 5% frequency criterion before subjecting them to an evaluation of their potential to discriminate individuals with respect to the target latent trait. For Example-item 1 and Example-item 2, there were no distractors with choice frequencies below 5%.

1.1.2. The Point-Biserial Correlation

Several indexes have been suggested that connect test performance (i.e., ability estimates) with distractor choice. Perhaps, the most popular index is the point-biserial correlation PB_D (Gierl et al. 2017; Attali and Fraenkel 2000) that contrasts test performance between participants who chose distractor D with the participants who did not choose D (i.e., the participants who chose either the correct solution or one of the other distractors):

$$PB_D = \frac{M_D - M}{S}\sqrt{\frac{P_D}{1 - P_D}}. \tag{1}$$

M_D is the average performance of participants who chose D, M is the average performance of all participants, S is the standard deviation of the performance of all participants, and P_D is the proportion of participants who selected D. Well-functioning distractors show a negative PB_D (Attali and Fraenkel 2000). However, Attali and Fraenkel (2000) pointed out that the groups of participants contrasted by PB_D do not yield the relevant information for developers in every situation. For example, a positive PB_D can be found for rather difficult items even when the average score of participants choosing D is substantially lower than the average score of participants solving the item (i.e., *M* is also affected by participants who chose one of the other distractors). Hence, they suggest an alternative index PB_{DC} that contrasts the group choosing *D* only with the group who solved the item:

$$PB_{DC} = \frac{M_D - M_{DC}}{S_{DC}} \sqrt{\frac{P_D}{P_C}}. \tag{2}$$

M_{DC} is the average sum correct score of the participants who either chose *D* or the correct solution *C*, S_{DC} is the standard deviation of the sum correct score of the group choosing either *D* or *C*, and P_C is the proportion of participants choosing the correct solution. It is clear that this index provides better contrast between distractor choice and item solution in terms of ability. However, both contrasts are not informative for the aim of detecting distractors with discriminatory power with respect to ability because this would require a contrast between participants choosing distractor *D* and participants choosing any other distractor. For Example-item 1 and Example-item 2, there were on average no differences observable for both PB_D and PB_{DC}. This is indeed expected given that both indices focus on a different aspect of discrimination as compared to the distractor discrimination parameters in nested logit models (see Table 1).

However, a corresponding variant of the point-biserial correlation is possible and could be calculated for each distractor, but this approach has two disadvantages: (a) Given that only participants who did not solve the item would be contrasted, such an index would be more prone to lacking empirical substance (i.e., very small group sizes for some of the distractor contrasts), and (b) looking at as many effect sizes as there are distractors in an item is expected to suffer from cumulative type-I-error (i.e., selecting an item for its discriminatory distractors when the true model behind the item has zero to negligible distractor discrimination). Consequently, we propose a different approach in Section 1.2 that circumvents these issues and relies on effect sizes at the item level.

1.1.3. Trace Line Plots and χ^2 Statistics

An alternative index connecting the ability with distractor choice extends a graphical tool labeled option characteristic curve by Levine and Drasgow (1983)—also known as option trace lines as it was labeled by Wainer (1989) or simply trace line plot (Gierl et al. 2017). For instance, choice frequency is plotted as a function of ability groups based on raw scores and the item options (i.e., the correct solution and each of the distractors). Figure 1 displays two trace line plots for Example-item 1 and Example-item 2. Clearly, in both plots, frequency of choosing the correct option was a positive function of ability (we used five ability groups here) with nearly the same trace line. For the distractors on the left side, it can be seen that they possessed varying attractiveness for the participants. Furthermore, distractor choice was a monotonically decreasing function of ability for all distractors. Haladyna and Downing (1993) suggested a χ^2_D statistic for each distractor to test whether distractor choice frequencies follow a uniform distribution across ability groups.

However, this statistic again focuses on each distractor in separation from the others and does not reveal anything about an interaction effect between distractor choice and ability group, which would be the crucial characteristic of the discriminatory power of distractors. This is also highlighted by Cohen's ω results based on the Haladyna-Downing approach (explicitly labelled ω_D to distinguish it from the effect size measure ω_G that will be introduced below) for Example-Item 1 and Example-Item 2 (see Table 1). On average ω_D was comparable in this case. One might argue that the variation of ω_D across distractors could be sensitive for the discriminatory power of distractors, but using a dispersion index (e.g., *SD* of ω_D across distractors) would yield a measure with a rather non-intuitive metric (as compared to commonly used effect size metrics). Thus, in this work, we aim at a direct effect size quantification of the discriminatory power of distractors. In addition, there were no intersections of the trace lines for the distractors between the ability groups in the left plot of Figure 1 (Example-Item 1). The χ^2_D statistic, however, would nonetheless be significant for all distractors in this plot (see also the large ω_D values in Table 1). In this vein, it has been conjectured by Garcia-Perez (2014) that non-monotonic empirical trace lines are required (such as those ones depicted for Example-Item 2 in the right plot of Figure 1) to allow effective modeling by polytomous IRT models. Hence, effect sizes are needed, which are sensitive for the detection of distractor trace lines that display distractor-ability interaction effects. In this work, we will use an effect size based on the χ^2 statistic using ability groups (as shown in Figure 1). In this approach, however, all distractors are considered (i.e., participants solving an item will be discarded from analysis); for a comparable implementation see Levine and Drasgow (1983). However, they used a less intuitive metric as the one that we will introduce in Section 1.2.

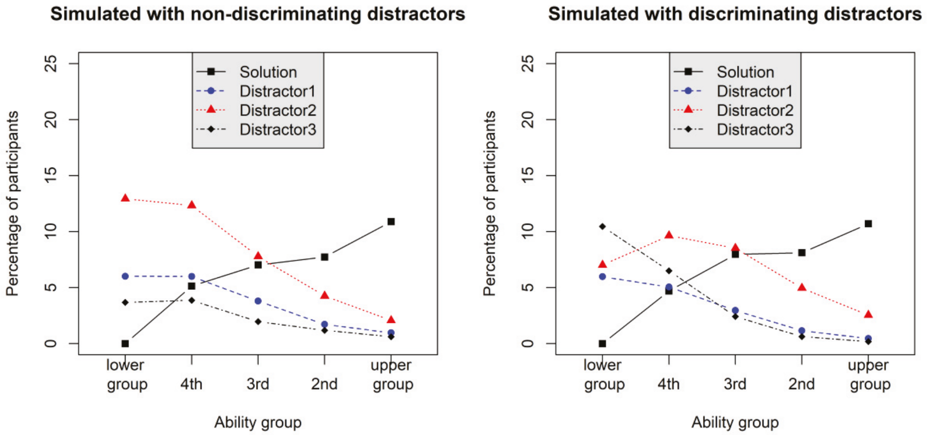


Figure 1. Trace line plots (Gierl et al. 2017; Wainer 1989) of two simulated items (*N* = 10,000 each) when distractors were simulated as non-discriminating in a 2PL (left plot) and when distractors were simulated with moderate NRM (nominal response model) discrimination in a 2PNL.

1.1.4. Rising Selection Ratios

Love (1997) suggested the criterion of rising selection ratios as a basic property of multiple-choice tests subjected to polytomous scoring. This criterion implies that the odds for choosing the correct option vs. distractor *D* is a monotone increasing function of ability. It is noteworthy, that this criterion does not require relative frequency of choosing *D* to be a monotone decreasing function of ability, because the probability of choosing the correct option relative to choosing *D* (i.e., the criterion of rising selection ratios) can be fulfilled with choice frequencies of *D* being a non-monotone function of ability [see Revuelta (2005) for applying this criterion to the 3PL and various polytomous IRT models]. Hence, primarily, the criterion of rising selection ratios is in accordance with modeling accuracy in the first place as a function of ability as it is put forth in nested logit models. At the same time, this criterion allows for interactions between ability and distractor choice behavior, due to the non-monotonicity of distractor choice in ability. Love (1997) suggested to use Goodman and Kruskal's γ coefficient (Goodman and Kruskal 1979) between test performance calculated by the sum total scores on all the other items (i.e., not the item under consideration for testing rising selection ratios) and the probability estimated from a $2 \times J$ contingency table with *J* as the number of possible test performance scores. The two rows include the frequencies for choosing the correct option and the frequencies for choosing *D* and the probability for choosing probability is then calculated by dividing the entries in the correct-option row by the respective column sums (Love 1997). However, this approach to evaluate data for rising selection ratios does not tap into potential interaction effects between ability and distractor choice. This is illustrated by the findings in Table 1. Goodman-Kruskal γ was found to be 1 for every distractor in Example-item 1 and Example-item 2, and hence, was not sensitive to the difference between these items in terms of distractor discrimination.

1.2. Effect Sizes for the Detection of Discriminatory Distractors

1.2.1. Cohen's ω Based on Ability Groups and Distractor Choice

We suggest Cohen's ω effect size based on the χ^2 derived from a contingency table in which the rows represent the distractors and the columns represent ability groups. Hence, this effect size is analogous to the above-mentioned approach used by Levine and Drasgow (1983), but it has a normed range that is easier to interpret. They also scale the χ^2 statistic for better interpretability. In particular, the χ^2 should be independent of the number of participants who did not solve the item under consideration because the raw χ^2 statistic would clearly depend on item difficulty otherwise (Levine and Drasgow 1983). Cohen's ω (Cohen 1992) can be calculated by

$$\omega_G = \sqrt{\frac{\chi_G^2}{\sum_{i=1}^{k} N_{D_k}}}. \tag{3}$$

G is the number of ability groups (e.g., as built by quantiles), χ_G^2 is the χ^2 statistic based on the G ability groups and all K distractors, and $\sum_{i=1}^{k} N_{D_k}$ is the number of all participants who did not solve the item under consideration. In this study, we will examine Cohen's well-known interpretation guideline for ω_G (Cohen 1992). Specifically, we will use 0.10 (small effect size), 0.30 (medium effect size), and 0.50 (large effect size) as cut-offs for the detection of items with discriminatory distractor sets. Example-Item 1 had ω_2 = 0.02 and ω_5 = 0.04, whereas, Example-Item 2 had ω_2 = 0.24 and ω_5 = 0.34. This illustrates that the discriminatory power of distractors could potentially be detected with this variant of Cohen's ω.

1.2.2. Canonical Correlation Based on Ability and Distractor Choice

Coefficient η has been suggested by Haladyna (2004) to be indicative of discriminatory power of item distractors. Accordingly, distractors with comparable choice means (implying a rather small η coefficient) render an item potentially less suitable for polytomous scoring as compared to an item with varying choice means. However, for reasons of a better conceptual fit, we will shift away from the η coefficient to the canonical correlation coefficient. The canonical correlation coefficient is known to be mathematically identical with coefficient η when one set of variables comprises of a number of binary indicator variables (i.e., the dummy variables also used for the calculation of η) and the other set includes only one continuous variable (Klecka 1980). However, the canonical correlation does not make the distinction between dependent and independent variable as it is the case for the η coefficient. The calculation of η is well aligned with the idea of mean comparisons of a continuous dependent variable (i.e., ability estimates) between groups that are defined by a categorical independent variable (i.e., distractor choices). However, in nested logit models, the relationship between ability and distractor choice is modeled vice versa: Distractor choice is modeled as a function of ability. Hence, we argue in favor of the canonical correlation because it does not suffer from this conceptual confusion, while simultaneously maintaining its potential for the detection of item-wise distractor discrimination.

In the context of this work, the canonical correlation is based on two sets of variables: (a) A matrix $\mathbf{X}_1$, including k binary indicator variables with an entry of 1 in the kth column and vth row when person v chose distractor k and zero otherwise, and (b) a vector $\mathbf{x}_2$ that includes the total scores for all participants who did not solve the item under consideration. Then, $\mathbf{r}_{12}$ is the column vector, including the correlations between each column from $\mathbf{X}_1$ and $\mathbf{x}_2$, and $\mathbf{H}_1$ is the Cholesky decomposition (Harville 2008) of the correlation matrix between all binary indicator variables in $\mathbf{X}_1$. With these terms in mind, the canonical correlation can be expressed as

$$R_{CC} = d_{11}, \tag{4}$$

with d_{11} is the only element from the **D** matrix resulting from a singular value composition (Harville 2008) of $\mathbf{W} = \mathbf{r}'_{12}\mathbf{H}_1^{-1}$. For the canonical correlation the same cut-offs for the detection of items with discriminatory distractors are suggested as it was the case above for Cohen's ω_G (small—0.10; medium—0.30; and large—0.50). Example-Item 1 had R_{CC} = 0.01, whereas, Example-Item 2 had R_{CC} = 0.34. This illustrates that the discriminatory power of distractors could also be detected by means of R_{CC}.

1.3. Aim of the Current Study

In the current work, a thorough simulation study was undertaken to examine the potential of Cohen's ω and R_{CC} (as outlined above) to detect items for their potential to discriminate individuals with respect to their latent trait based on distractor choice behavior. To this aim, we first simulated conditions in which distractors did not possess discriminatory power with respect to the latent trait to assess the type-I-error of the used statistical indices (i.e., effect sizes passing the effect size threshold, when the population model did not include discriminatory distractors). Second, we simulated conditions based on a population model with discriminatory distractors to examine the power to detect items that are suitable for nested logit modeling. A final aim of this work is to illustrate the suggested item-analytical strategy by means of the data taken from Myszkowski and Storme (2018).

2. Simulation Study

2.1. Method

2.1.1. Data Generating Model

The data were simulated according to a 2PNL (Suh and Bolt 2010) in which the probability that person j solves item i is modeled by the following logistic model:

$$P(x_{ij} = u|\theta_j) = \frac{1}{1+e^{-(\beta_i+\alpha_i\theta_j)}}. \tag{5}$$

u is the correct option, θ_j is the ability parameter, β_i is the item difficulty parameter, and α_i is the discrimination parameter. Then, in case that an item has not been solved the probability to choose distractor v among the set of the remaining m_i distractors is modeled by the nominal response model with intercept parameters ζ_{iv} and distractor discrimination parameters λ_{iv}:

$$P(x_{ij} = v|\theta_j) = \left[1 - P(x_{ij} = u\middle|\theta_j)\right]\left[\frac{e^{\zeta_{iv}+\lambda_{iv}\theta_j}}{\sum_{k=1}^{m_i} e^{\zeta_{ik}+\lambda_{ik}\theta_j}}\right]. \tag{6}$$

2.1.2. Facets of the Simulation Design

Several factors were manipulated to allow a thorough investigation of the usefulness to detect items with discriminatory distractors for nested logit modeling:

1. Sample size (three levels): N = 100; N = 200; and N = 500.
2. Number of items (three levels): I = 10; I = 20; and I = 50.
3. Number of distractors (two levels): D = 3; and D = 7.
4. 2-PL difficulty (three levels): Moderate [$\beta_i \sim U(-0.15, 0.15)$]; difficult [$\beta_i \sim U(-1.15, -0.85)$]; and very difficult [$\beta_i \sim U(-2.25, -1.85)$].
5. 2-PL discrimination (three levels): Low [$\alpha_i \sim U(0.25, 0.55)$]; moderate[$\alpha_i \sim U(0.85, 1.15)$]; and high [$\alpha_i \sim U(1.60, 1.90)$].
6. NRM discrimination parameters (four levels) are depicted in Table 2.

Table 2. NRM discrimination parameters used in the simulation study.

Distractor	Level 1—Zero	Level 2—Moderate	Level 3—High	Level 4—Very High
	3 distractors/7 distractors	3 distractors/7 distractors	3 distractors/7 distractors	3 distractors/7 distractors
λ_{i1}	0.00/0.00	−0.40/−1.20	−1.00/−3.00	−1.75/−5.25
λ_{i2}	0.00/0.00	0.00/−0.80	0.00/−2.00	0.00/−3.50
λ_{i3}	0.00/0.00	0.40/−0.40	1.00/−1.00	1.75/−1.75
λ_{i4}	-/0.00	-/0.00	-/0.00	-/0.00
λ_{i5}	-/0.00	-/0.40	-/1.00	-/1.75
λ_{i6}	-/0.00	-/0.80	-/2.00	-/3.50
λ_{i7}	-/0.00	-/1.20	-/3.00	-/5.25
NRM discrimination (step size)	0.00	0.40	1.00	1.75

See DeMars for a comparable approach to simulate items with discriminating distractors according to the NRM (DeMars 2003). NRM discrimination (step size): This is the step size between the consecutive λ_i parameters that can be used as a general indicator of NRM discrimination.

NRM intercepts ζ_{iv} were further sampled for all design cells from a $U(-1, 1)$ distribution. Further facets resulted from the used effect size threshold and the type of effect size (but these facets did not imply additionally generated datasets):

7. Effect size threshold (three levels): Small: Effect size > 0.10; moderate: Effect size > 0.30; and large: Effect size > 0.50.
8. Type of effect size (three levels): Cohen's ω based on two ability groups; Cohen's ω based on five ability groups; and the canonical correlation coefficient.

2.1.3. Dependent Variables

The main dependent variable was: (1) The proportion of effect sizes that were larger than the effect size threshold. In addition, we examined the following dependent variables related to the empirical substance of the simulated datasets: (2) the proportion of distractors with relative choice frequencies smaller than 5%; (3) the proportion of missing effect sizes; (4) the proportion of missing effect sizes resulting from too many distractors with relative choice frequencies smaller than 5%; and (5) the proportion of ability groups occurring in the simulated data. For example, a value of 0.99 for Cohen's ω based on five ability groups and in case of 10 simulated items implies that $0.99 \times 5 \times 10 = 4950$ groups were simulated out of 5000 possible groups.

2.1.4. Simulation Setup

All simulations and analysis were carried out by means of the statistical software R (R Core Team 2019). The simulation of the datasets was performed with the `simdata()` function included in the R package mirt (Chalmers 2012). The design for the dataset generation was based on crossing all facets of the simulation design (see 1. to 6. presented in Section 2.1.2). Hence, it was a sample-size × number-of-items × number-of-distractors × 2-PL-difficulty × 2-PL-discrimination × NRM-discrimination design with $3 \times 3 \times 2 \times 3 \times 3 \times 4 = 648$ cells. For each of these 648 cells, we generated 1000 datasets and aggregated the dependent variables across these datasets for each cell. All R code files and simulated data are available in the OSF repository for this work (https://osf.io/9tp8h/).

2.2. Simulation Results

2.2.1. Type-I-Error Results

Results with respect to type-I-error revealed a clear picture of findings. First, an effect size threshold of 0.10 appeared to be far too liberal regardless of any other design facet. In fact, for all cells in the design, the type-I-error rate for the 0.10 threshold was clearly above the conventional 0.05 level (see Figure 2). Second, the worst performance in terms of type-I-error rate was observed for Cohen's ω based on five ability groups. This effect size measure reached only acceptable type-I-error rates for very specific conditions (see Figure 2). For example, with three distractors and a 0.30 threshold, ω_5 was adequate only when the sample size was $N = 500$. Moreover, for seven distractors, a 0.30 threshold, and a sample size of $N = 500$ acceptable type-I-error rates were reached only for very difficult items. The best performance of Cohen's ω based on five ability groups was found for the three-distractor condition and a threshold of 0.50 (i.e., only type-I-error for moderately difficult items was too large). However, both other effect size measures (Cohen's ω based on two ability groups and the canonical correlation) yielded highly conservative type-I-error rates (i.e., type-I-error rates that are notably smaller than 0.05) when the effect size threshold was 0.50 regardless of any other design facet. Moreover, Cohen's ω based on two ability groups and the canonical correlation coefficient yielded acceptable to conservative type-I-error rates with three distractors and a threshold of 0.30 when 2PL-difficulty was at least difficult. The same was observed for these two effect size measures for seven distractors, but only when the sample size was at least $N = 200$ (see Figure 2). Finally, we found that 2PL-difficulty was inversely related to type-I-error rates as in several simulated conditions moderate 2PL-difficulty resulted in the highest type-I-error rate (see Figure 2), whereas, the level of 2PL-discrimination and the number of items did not show any specific relationship with a type-I-error rate (see Appendix A).

Based on these findings, we refrain from any power examinations for conditions with an effect size threshold of 0.10. It is further noteworthy that for all other effect size thresholds Cohen's ω_2 (85%) and R_{CC} (84%) comparable numbers of cells with acceptable type-I-error rates resulted, whereas, Cohen's ω_5 displayed acceptable type-I-error rates only for 48% of the simulated cells. Thus, Cohen's ω_5 appeared to have only a very narrow range of scenarios in which this statistic is advisable for the detection of discriminatory distractors. Cohen's ω_2 and R_{CC}, however, were found to function comparably well (see also Table 3).

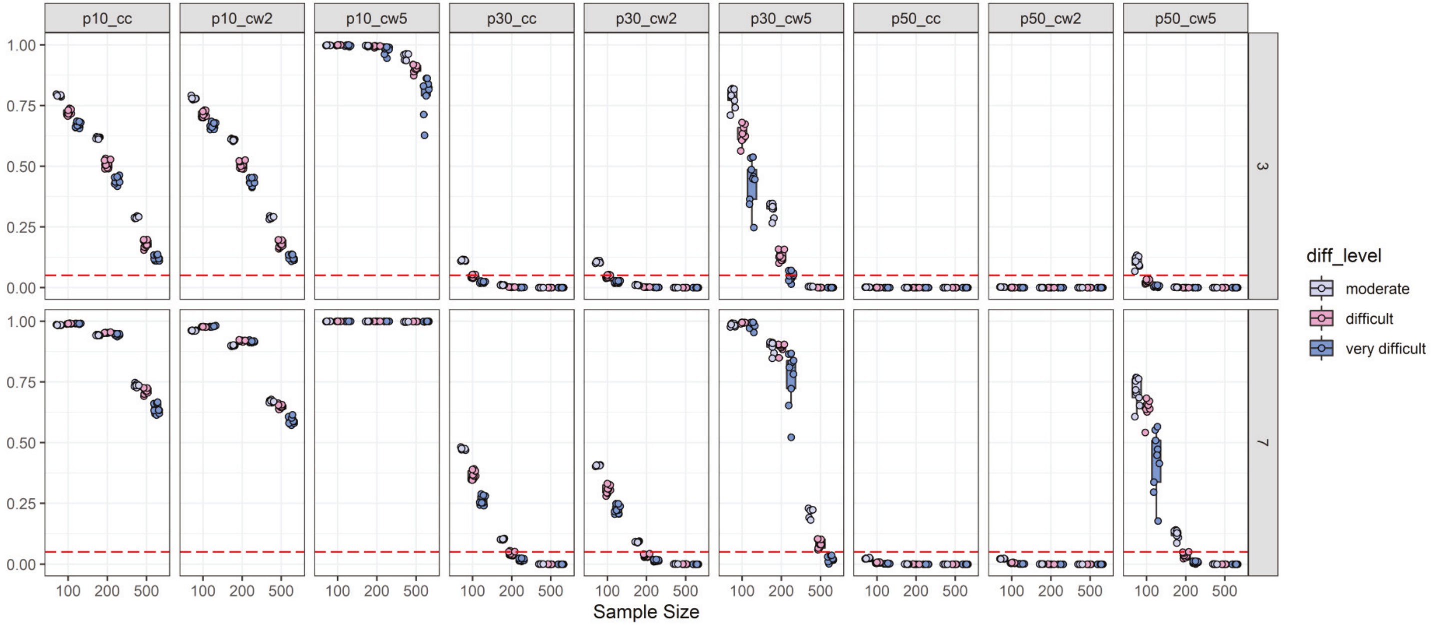


Figure 2. 2PL-difficulty split of type-I-error analysis: Depiction of the type-I-error rate (y-axis) as a function of sample size (x-axis), number of distractors (three distractors = top-row; seven distractors = bottom-row), 2PL-difficulty (diff_level: Moderate vs. difficult vs. very difficult), and effect size measures combined with effect size thresholds (see explanation) (p10_cc = canonical correlation with a 0.10 threshold; p10_cw2 = Cohen's ω based on two ability groups with a 0.10 threshold; p10_cw5 = Cohen's ω based on five ability groups with a 0.10 threshold; p30_cc = canonical correlation with a 0.30 threshold; p30_cw2 = Cohen's ω based on two ability groups with a 0.30 threshold; p30_cw5 = Cohen's ω based on five ability groups with a 0.30 threshold; p50_cc = canonical correlation with a 0.50 threshold; p50_cw2 = Cohen's ω based on two ability groups with a 0.50 threshold; p50_cw5 = Cohen's ω based on five ability groups with a 0.50 threshold). The horizontal red dashed line represents the target type-I-error rate of 0.05.

Table 3. Percentages of cells in the simulation design with adequate type-I-error rate and empirical power.

	Threshold = 0.30			Threshold = 0.50		
	R_{CC}	ω_2	ω_5	R_{CC}	ω_2	ω_5
Adequate type-I-error rate						
All	69	**70**	25	**100**	**100**	72
Adequate power						
NRM Discrimination = 0.40						
All	**23**	17	5	0	0	**7**
$M(\gamma) > 0.30$	**17**	3	4	-	-	**6**
$M(PB_{DC}) < -0.30$	**16**	3	4	-	-	**3**
NRM Discrimination = 1.00						
All	**61**	45	23	**35**	24	23
$M(\gamma) > 0.30$	**39**	31	16	**25**	17	14
$M(PB_{DC}) < -0.30$	**41**	30	16	**24**	15	15
NRM Discrimination = 1.75						
All	**65**	62	26	**65**	41	54
$M(\gamma) > 0.30$	**38**	**38**	14	**41**	25	36
$M(PB_{DC}) < -0.30$	40	**41**	15	**44**	24	40

Percentages are rounded to integers. The threshold-specific best-performing statistic is highlighted in bold. When two effect sizes performed equally well, both were highlighted. The total number of cells for each of the respective levels of NRM discrimination was 162 (please note that the cells for checking type-I-error rates had zero NRM discrimination). $M(\gamma) > 0.30$: In addition to adequate empirical power (≥0.80) the boundary condition that the average of Goodman-Kruskal's γ to check rising selection ratios had to be greater than 0.30. $M(MP_{DC}) < -0.30$: In addition to adequate empirical power the boundary condition that PB_{DC} to check discrimination between item solvers and participants who chose a certain distractor had to be smaller than −0.30. Frequencies in bold font refer to the best performing effect sizes under the respective threshold conditions.

2.2.2. Power Results

Prior to power analysis, all results for conditions that yielded unacceptable type-I-error rates were removed. Given that effect size measures are studied for their potential usefulness in the context of test-development, it is unlikely that other important item statistics, such as Goodman and Kruskal's γ to test for rising selection ratios or PB_{DC} would be ignored. Hence, we checked all conditions that had both acceptable type-I-error and sufficient power (i.e., power ≥ 0.80) for their power under additional boundary conditions. First, the power of effect size measures was reevaluated under the additional condition that the average γ is greater than 0.30. Second, another reevaluation of the power of effect size measures took PB_{DC} as a boundary condition into account. Here we tested the additional condition that the average PB_{DC} had to be smaller than −0.30. Table 3 displays the percentages of design cells with adequate empirical power with and without boundary conditions.

Across various conditions, the percentage of design cells with adequate power was highest for the canonical correlation (see Table 3). The only exception to this pattern was the moderate NRM discrimination condition paired with an effect size threshold of 0.50. However, in these conditions the best-performing statistic was ω_5 and adequate power was only achieved for less than 10% of the design cells, which was still surpassed by R_{CC} paired with a 0.30 threshold (see Table 3). Results indicated further, as expected, a positive relationship between NRM discrimination and empirical power. That is, the higher the NRM discrimination was in the data-generating model; the higher was the percentage of design cells with adequate power to detect discriminatory distractors (see Table 3). This pattern was rather robust across effect size thresholds and the used effect sizes. Restricting the findings to R_{CC} as the overall best-performing statistic, however, revealed that power gains from *high* to *very high* NRM conditions were negligible (even non-existent when boundary conditions were taken into account) with a 0.30 threshold. Comparing further the R_{CC} results between the 0.30 and the 0.50 threshold, independent of NRM discrimination, suggested that the power advantage of the lower

0.30 threshold as compared to the 0.50 threshold vanishes for *very high* NRM discrimination. The differences between power analyses without and with boundary conditions increased with the level of NRM discrimination. Generally, the overall impression of empirical power results was supported regardless of the presence of additional boundary conditions.

In Figure 3, the power simulation results are split according to NRM discrimination conditions between panels and according to the number of items. We found that empirical power was positively related to the number of items and the number of distractors, indicating that more items and more distractors increase empirical power to detect informative distractors. The simulation results might give further the impression that for most of the conditions, sample size was negatively related to empirical power. However, when comparing Figure 3 with Figure 5, it becomes clear that this impression occurs, due to the influence of the low 2PL discrimination conditions on sample-size-specific empirical power (i.e., there seems to be an interaction between sample size and 2PL discrimination level). Otherwise, some conditions revealed a positive relationship between empirical power and sample size. For example, for high NRM discrimination, three distractors, at least 20 items, and for a threshold of 0.30, empirical power was a positive function of sample size for both the canonical correlation coefficient and Cohen's ω based on two ability groups (see Figure 3). For these conditions, empirical power also surpassed the 0.80 level for sufficient power. Moreover, detection of moderately discriminate distractors was possible by means of the canonical correlation, but only with seven distractors, at least 20 items and a threshold of 0.30. Under the same conditions, Cohen's ω needed at least 50 items for sufficient power. Detection of discriminatory distractors with three distractor items required at least a high NRM discrimination with again the best findings for the canonical correlation that required at least 20 items for adequate power (as compared to at least 50 items for Cohen's ω).

In Figure 4, the boxes of the boxplots are depicted in different colors depending on 2PL difficulty. While recognizing that item easiness is negatively associated with the available data for participants choosing one of the distractors, this plot also suggests a picture in line with the idea that the effect sizes are subject to an upward bias. Here again, we suggest a cautious interpretation, because this impression was driven by low 2PL-discrimination conditions (see Figure 5) that were presented among the other findings for the varying difficulty conditions. Again, under these various 2PL difficulty conditions, the canonical correlation combined with a 0.30 threshold displayed the best findings with respect to empirical power across various conditions. The exceptions from this pattern can be inferred from Figure 5 to be caused by conditions in which 2PL discrimination was low (see the skyblue boxplots in the subplots for p30_cc). Hence, one might conclude that after a check of item-scale correlations, the canonical correlation combined with a 0.30 threshold seems to be the best choice for the task of pre-selecting items with discriminatory distractors from a pilot study (please note that this conclusion also takes type-I-error into account, because power was only examined for conditions with acceptable type-I-error rates). The detailed power findings presented in Figures 3–5 replicated well under additional boundary conditions as it was the case for the results aggregated in Table 3. Appendix B provides detailed figures of power findings under boundary conditions.

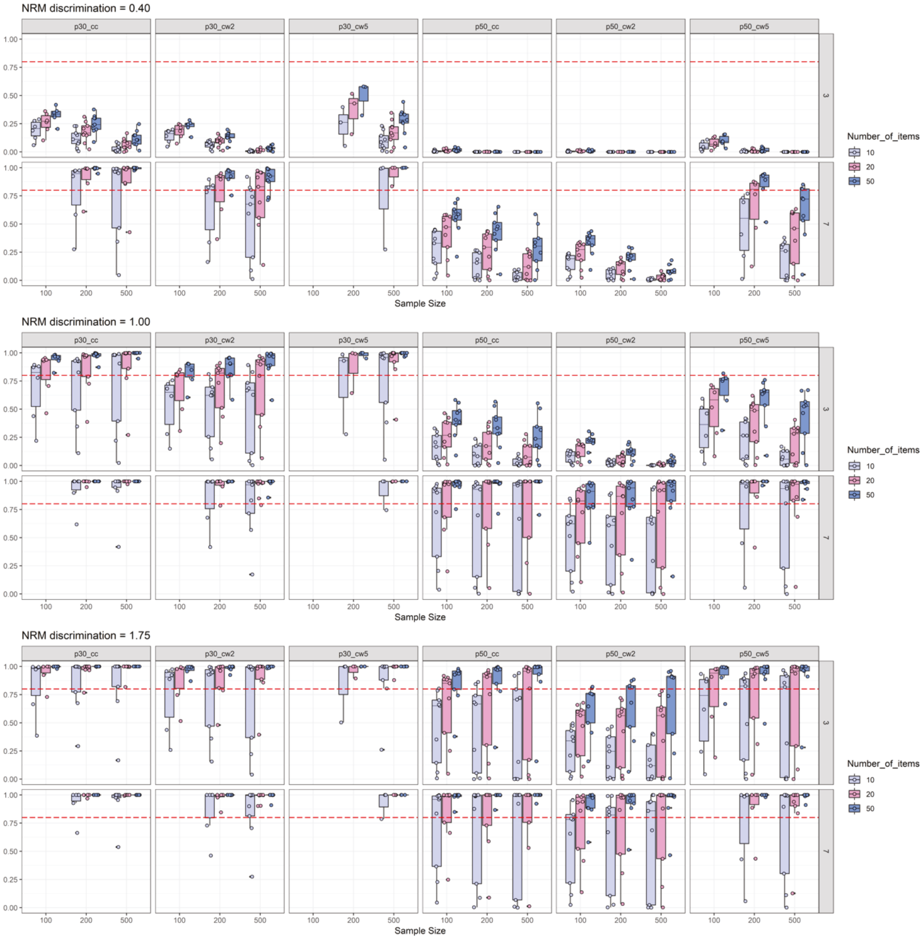


Figure 3. Number-of-items split of empirical power analysis: Depiction of the empirical power (y-axis) as a function of NRM discrimination (0.40 = top-row; 1.00 = middle-row; and 1.75 = bottom-row), sample size (x-axis), number of distractors (three distractors = top-row in each sub-plot; seven distractors = bottom-row in each sub-plot), number of items, and effect size measures combined with effect size thresholds. The horizontal red dashed line represents the target power level of 0.80. For more explanations, see Figure 2.

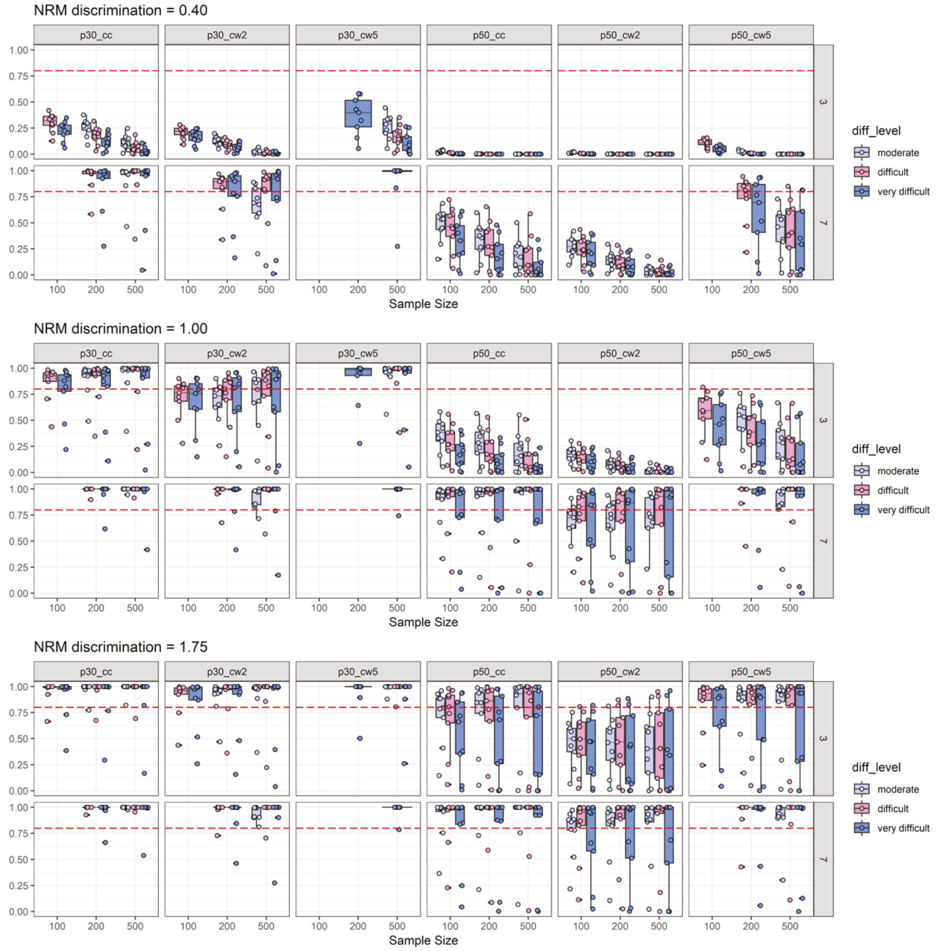


Figure 4. 2PL-difficulty split of empirical power analysis: Depiction of the empirical power (y-axis) as a function of NRM discrimination, sample size (x-axis), number of distractors, 2PL-difficulty (diff_level: Moderate vs. difficult vs. very difficult), and effect size measures combined with effect size thresholds. The horizontal red dashed line represents the target power level of 0.80. For more explanations, see Figures 2 and 3.

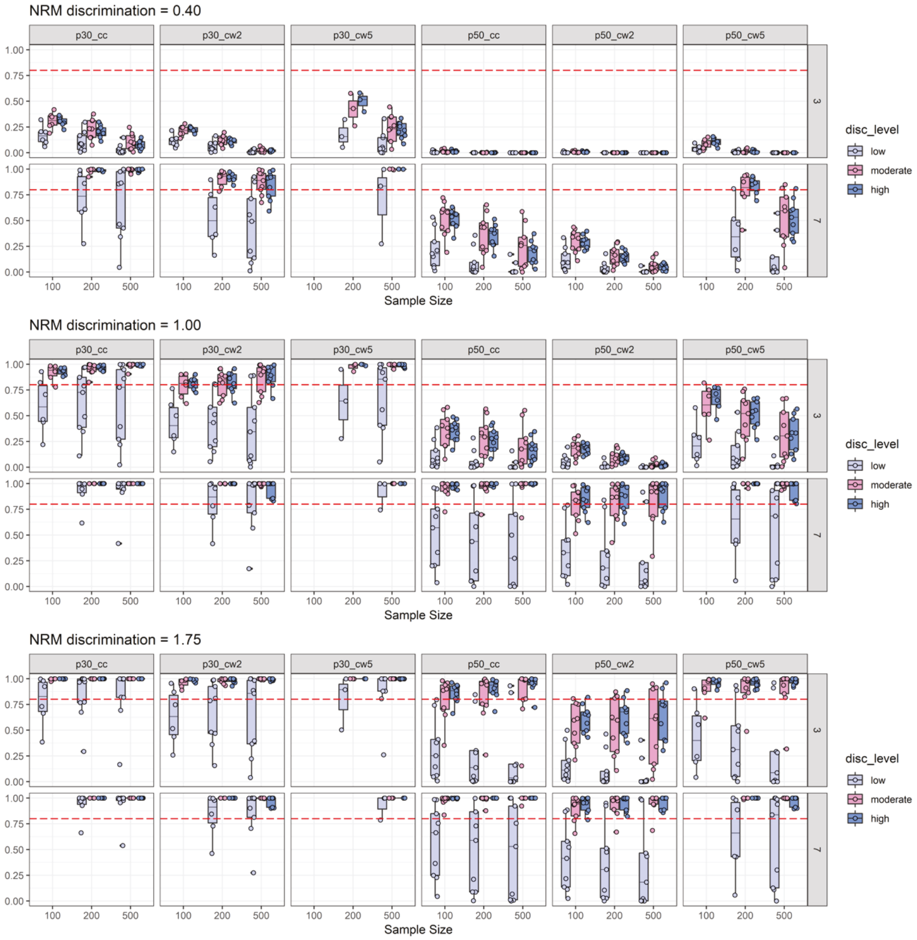


Figure 5. 2PL-discrimination split of empirical power analysis: Depiction of the empirical power (y-axis) as a function of NRM discrimination, sample size (x-axis), number of distractors, 2PL-discrimination (disc_level: Low vs. moderate vs. high), and effect size measures combined with effect size thresholds. The horizontal red dashed line represents the target power level of 0.80. For more explanations, see Figures 2 and 3.

2.2.3. Empirical Substance Examination

We further examined the empirical substance for each of the simulation conditions. First, the percentage of distractors with relative choice frequencies smaller than 5% increased with NRM discrimination (9.62%, 10.34%, 15.56%, and 21.33% for NRM discrimination levels of 0, 0.40, 1.00, and 1.75, respectively). The percentage of missing effect sizes, due to distractor choice frequencies $< 5\%$, however, was rather small in conditions (the maximum was $< 1\%$ for zero and moderate NRM discrimination; and maximally 1.67% or 3.96% for NRM discrimination levels of 1.00 and 1.75, respectively). The overall percentage of missing effect sizes was then examined, and across all 648 cells of the simulation design, we found 17 cells with percentages of missing effect sizes larger than 1%. An amount of 1% of missing effect sizes implies, for example, that for 1000 replications and 10 items the number of missing effect sizes would be 100. The largest percentage of missing effect sizes was found to be 4% for the condition with $N = 500$, $I = 10$, $D = 7$, moderate 2PL-difficulty, high 2PL-discrimination, and the highest level of NRM discrimination. All 17 cells had in common that $D = 7$, 2PL-difficulty was moderate, 2PL-discrimination was high, and that NRM discrimination was at least high. For all other design cells, the percentage of missing effect sizes was below 1%, and for most of the cells, this percentage was zero or negligible. The minimal proportion of occurring group sizes in the data was 99.99% for ω_2 in all conditions and 79.95% for ω_5 in all conditions. Hence, ω_5 was affected the most by empirical substance loss, which in turn might explain its inferior performance in the simulation study.

2.2.4. Discussion of Simulation Study Findings

In this simulation study, we thoroughly investigated the type-I-error rates and empirical power of R_{CC}, ω_2, and ω_5 effect sizes to detect the discriminatory power of distractors. The power examination was also carried out under additional boundary conditions defined by effect sizes with a focus on solution behavior (i.e., γ and PB_{DC}). The simulation was further flanked by an empirical substance investigation to reveal the amount of information loss when, for example, distractors are chosen by less than 5% of the participants or creation of ability groups did not result in the target number of groups. The aim of this simulation was twofold: (a) We wanted to identify the best-performing effect size for the detection of discriminatory distractors, and (b) we wanted to explore potential factors that influence type-I-error and empirical power.

Results suggested that R_{CC} and ω_2 yielded comparable performance with respect to type-I-error. R_{CC} and ω_2 displayed acceptable type-I-error for a far greater variety of simulated conditions as compared to ω_5. Hence, ω_5 was found to be clearly limited in its range of application. In terms of empirical power, however, it was found that R_{CC} clearly outperformed ω_2 in most of the simulated conditions with few design cells in which R_{CC} and ω_2 performed comparably well. In relation to this, it is further important that using R_{CC} in combination with a 0.30 threshold yielded better empirical power findings in conditions with moderate or high NRM discrimination. For very high NRM discrimination conditions R_{CC} combined with a 0.30 threshold and R_{CC} combined with a 0.50 threshold were found to be comparable with respect to empirical power findings. Thus, for a wide range of simulated conditions in this study, R_{CC} in combination with a 0.30 threshold would be the best choice.

A more fine-grained analysis of influencing factors on type-I-error and empirical power revealed that it is not generally recommended to use sample sizes of $N = 100$ for the detection of discriminatory distractors. Based on type-I-error findings, the sample size should be at least $N = 200$ when items include three distractors. When items include seven distractors, however, sample sizes below $N = 500$ cannot be recommended without further considerations, due to unacceptable type-I-error rates. In relation to this, it needs to be noted that type-I-error rates for a threshold of 0.50 were acceptable for all conditions when R_{CC} as the generally best-performing effect size measure is used (i.e., ω_2 also had acceptable type-I-error rates for all conditions with a 0.50 threshold, but did not perform on par with R_{CC} with respect to power). However, in terms of empirical power, it is important to take further into account that 2PL-discrimination needs to be at least moderate and NRM discrimination had to be very high to yield largely acceptable detection power (with better findings for seven distractor items). When NRM discrimination is only high, detection of discriminatory distractors was only feasible for seven distractor items when at the same time 2PL-discrimination was at least moderate. Empirical power, with a 0.50 threshold to detect items with moderate NRM discrimination, was found to be unacceptable.

Importantly, the presented findings on the detection of ability-related discriminatory distractors suggest that there is no simple rule to increase the power analogous to experimental study planning (e.g., the more participants, the higher the power to detect a certain assumed mean difference between experimental groups). Simply raising sample size or the number of items (or even both) did not generally increase empirical power in the simulation. For example, with increasing sample sizes or increasing difficulty, it was found that variation in empirical power increased for low 2PL-discrimination conditions (see Figures 3–5). Hence, 2PL-discrimination seems to be a precondition before power follows the commonly known "the-more-the-better" rule of thumb. In light of these specificities of the simulation findings, we, thus, recommend researchers to formulate their expectations (e.g., based on previous empirical studies) about a potential item pool with respect to important parameters that were found to be influential in this study, such as 2PL item difficulty (items should be difficult or very difficult) and 2PL item discrimination (items should have at least moderate 2PL discrimination) and run their own customized simulation to guide scale development. For example, scale development must be efficient sometimes, and it could be the case that only N = 150 participants are available for a first examination of the ability-related discriminatory power of distractors. Then, with the simulation code of this study as a starting point (the code is openly available at https://osf.io/9tp8h/), it is possible to explore different thresholds (i.e., also thresholds between 0.30 and 0.50) with respect to type-I-error and empirical power in combination with other characteristics (e.g., number of items or number of distractors) to choose the best design for scale development. Most likely, a design with $N < 100$ will not be applicable which prevents usage of the suggested approach at very early stages of scale development in which items are tested for clear instructions or the wordings of item content (Johanson and Brooks 2010).

3. Empirical Illustration

3.1. Method

3.1.1. Dataset

The studied dataset was taken from Myszkowski and Storme (2018). This dataset includes N = 499 participants from a French business school (undergraduates; 285 females and 214 males; age: M = 20.70, SD = 0.93). All participants worked on the last series of Raven's Standard Progressive Matrices (SPM-LS) without any imposed time limit. The instructions further encouraged the participants to provide a response even when they were unsure about the correct solution (Myszkowski and Storme 2018). Thus, no missing data are present in this dataset. The SPM-LS consists of twelve items with seven distractors each. Hence, this dataset closely mimics the design in the simulation study above with N = 500, I = 10, and D = 7.

3.1.2. Analytical Strategy

This empirical illustration will apply the effect size measures introduced in this work. Based on the findings of the simulation study, we calculated R_{CC} as effect size with a threshold of 0.30 because it displayed acceptable type-I-error rates and reasonable empirical power under comparable conditions as given for the given dataset in the simulation study above. In addition, the number of distractors with relative choice frequencies < 0.05, PB_{DC} and γ were calculated. Finally, we re-estimated the 2PL-parameters to facilitate interpretation of the findings in connection with the simulation study presented above.

3.2. Results and Discussion

The findings presented in Table 4 revealed that for items 1 to 5 distractor choice frequencies were too sparse to use the R_{CC}. As expected, this sparsity of distractor frequencies was associated with 2PL-difficulty estimates. These five items were indeed among the easiest items according to the estimates in Table 4. Moreover, the estimates, in particular those for items 1 to 5, were much higher as compared to the difficulties simulated above. In fact, only items 10 to 12 were found to be in the range of simulated 2PL-difficulty values used above. The values for items 8 and 9 were closer to the moderate difficulty level used in the simulation, whereas, the estimates for items 6 and 7 were clearly easier. The 2PL-discrimination estimates, however, were inside the range of the simulation study and were even higher for several items. The latter observation is particularly important, because even for the detection of moderate NRM discrimination it was found that R_{CC} had adequate power levels with seven distractors and a sample size of N = 500 (which are conditions resembling the Myszkowski-Storme dataset). Given that 2PL-discrimination was identified in the simulation as an important influencing factor on the detection power, one could reasonably assume that higher 2PL-discrimination can compensate for lower 2PL-difficulty as compared to the used simulation setup. This reasoning applies particularly to item 6, which was found to have a much larger 2PL-discrimination parameter estimate as compared to the values used in the simulation (and a much lower difficulty estimate). Nonetheless, caution is needed when interpreting these findings with parameter estimates outside the simulated values.

Table 4. Distractor choice frequency, 2PL-parameter estimates, and distractor effect size measure findings on the Myszkowski-Storme dataset.

Item	Number of Distractors with Relative Choice Frequency < 0.05	2PL-Difficulty	2PL-Discrimination	R_{CC}	$M(PB_{DC})$ [2]	$M(\gamma)$ [3]
Item 1	6	1.32	0.85	NA [1]	NA [1]	NA [1]
Item 2	7	3.56	2.01	NA [1]	NA [1]	NA [1]
Item 3	6	2.07	1.69	NA [1]	NA [1]	NA [1]
Item 4	6	4.11	4.10	NA [1]	NA [1]	NA [1]
Item 5	7	5.51	4.97	NA [1]	NA [1]	NA [1]
Item 6	5	2.13	2.38	0.46	−0.38	0.48
Item 7	4	1.23	1.55	0.28	−0.36	0.40
Item 8	1	0.50	1.61	0.34	−0.47	0.65
Item 9	3	0.40	1.27	0.34	−0.39	0.37
Item 10	1	−0.70	2.20	0.36	−0.61	0.77
Item 11	1	−0.82	1.51	0.21	−0.48	0.21
Item 12	1	−0.91	1.14	0.31	−0.43	0.23

[1] Distractor effect size measures were not calculated when the number of distractors with relative choice frequency < 0.05 exceeded a value of five (i.e., when only one distractor remained for analysis). [2] The average of all PB_{DC} values for all available distractors of an item is reported—the lower the average PB_{DC}, the better. [3] The average of all γ values for all available distractors of an item is reported—the higher the average γ, the better. The R code to reproduce the findings in this table can be found in Appendix C.

Analysis of distractor effect size measures revealed five items (6, 8, 9, 10, and 12) with canonical correlation coefficients > 0.30. Hence, we would suggest that the items flagged for distractors with discriminatory power by means of the canonical correlation are most likely the ones driving the reliability gain at the low-ability range, as reported by Myszkowski and Storme (2018). Moreover, these items are expected to fit a nested logit modeling approach well in a larger sample.

To secure these observations, we further calculated the PB_{DC} for items 6 to 12 to examine if the correct solution was associated with higher ability levels as compared to choosing one of the distractors. In addition, γ was used to check the rising selection ratio property. Table 4 displays the average PB_{DC} across all distractors for each of the items. These values ranged from moderate to large effect sizes implying rather well-functioning distractors in this regard. Importantly, some items displayed comparable R_{CC} values (items 9 and 10), but clearly varying PB_{DC} values. This highlights the importance to study ability-related discrimination of distractors and discrimination with respect to solution behavior at the same time. PB_{DC} and also γ focus on solution behavior, but R_{CC} focus on distractors that are more often chosen by participants with higher ability levels as compared to other distractors (i.e., not in comparison to the correct solution). These aspects of discrimination are not necessarily expected to covary. This is further illustrated by the R_{CC} and PB_{DC} findings for item 11, which had the lowest R_{CC} value, but the second strongest PB_{DC} (see Table 4).

The average γ findings were in the range from small to large, with an average γ smaller than 0.30 for items 11 and 12. This highlights that for an item not all boundary conditions might be fulfilled (in these cases average PB_{DC} was below −0.30, but γ was not larger than 0.30). To decide if this pattern is problematic for the detection of ability-related discriminatory distractors, new simulations to understand the interplay of PB_{DC} and γ as boundary conditions are clearly needed. Hence, results for items 11 and 12 should also be treated with caution. These findings largely support the feasibility of nested logit modeling with its primary focus on solution behavior and distractor choices as additional information used for ability estimation.

4. Overall Discussion

In our study, we suggest an item-analysis procedure to detect items with potentially discriminating distractors that can be used for trait estimation in models, such as nested logit models which take both accuracy and distractor choices into account. We thoroughly examined the usefulness of different effect sizes for distractor discrimination by a simulation study, and illustrated our findings by an application to an empirical dataset with participants who worked on the short form of Raven's Progressive matrices test. As such, our analysis had a different focus as compared to traditional distractor analyses which are usually concerned with distractor choice frequency and variants of biserial correlations to evaluate distractor quality (Gierl et al. 2017; Haladyna and Downing 1993; Attali and Fraenkel 2000; Haladyna 2004). Instead, Cohen's ω was examined as an effect size that can potentially reveal interactions between ability groups and distractor choice frequencies, which are indicative of the discriminatory power of distractors for the trait under consideration. As a second effect size, the canonical correlation was studied as a measure for the detection of the discriminatory power of item distractors in terms of the latent trait variable. The simulation revealed that in contrast to Cohen's ω, the canonical correlation coefficient seems to be most promising for the task of detecting items with discriminatory distractors.

Limitations

This work is limited to the simulation conditions chosen. For example, Myszkowski and Storme (2018) highlight the importance of taking item guessing into account. That is, they suggest relying on the 3PNL instead of relying on the 2PNL as we used in the simulation. Likewise, the 4PNL was not studied here to reduce the complexity of the simulation design. We argue that for a starting point to understand mechanisms behind item selection based on distractor effect size measures, the design was already rather complex. Future studies are clearly required in this regard.

In the empirical illustration, the suggested effect sizes for distractor discrimination were flanked by the average PB_{DC}, and the average γ and this approach seems to be most promising. In particular, the PB_{DC} can further reveal if choosing the correct solution is more strongly related to the ability as compared to any other distractor. This is crucial for a model that puts solution behavior in the first place. Moreover, average γ ensures that the assumption of rising selection ratios holds for a set of candidate items which is further useful to guide item pre-selection. In our simulation, we found that using these two statistics as boundary conditions is useful, but for simplicity, these two item statistics were studied in isolation. Obviously, other more complex item-analysis strategies could be used in which also a combined cut-off for both statistics is used. It is further possible to consider scenarios in which item-selection is carried out in multiple consecutive steps, and the current work does not shed much light into the question which statistic should be consulted first (e.g., testing discrimination in the sense of item-scale correlations first, testing for rising selection ratios second, and screening for ability-related discriminatory power of distractors as a final step).

5. Conclusions

Overall, the potential of nested logit models to construct measures with higher measurement precision at lower ability levels by means of exactly the same items is highly attractive not only for cognitive ability constructs as intelligence, but especially for measures that need to be very short and efficient. However, the long tradition of, for instance, figural matrix items and available theories with respect to solution behavior, item-, and distractor generation principles seem to be a key requisite for the construction of such a measure. The canonical correlation as an effect size for distractor discrimination seems to be a promising statistical tool for pre-selecting useful items for scale construction in this regard.

In fact, the canonical correlation can be interpreted analogously to classical item-scale correlations (i.e., test developers can rely on a familiar 0.30 cut-off). Moreover, it fits the idea of item response models in which observable behavior (i.e., distractor choice) is modeled as a function of ability.

Author Contributions: Conceptualization, B.F.; Methodology, B.F.; Software, B.F.; Validation, B.F.; Investigation, B.F.; Resources, E.S.; Data Curation, B.F.; Writing-Original Draft Preparation, B.F., N.F., B.S.; Writing-Review & Editing, B.F., N.F., B.S., K.H., J.F., M.T.P., E.S.; Visualization, B.F., N.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

In this appendix, further simulation results on the type-I-error rate are displayed in Figures A1 and A2.

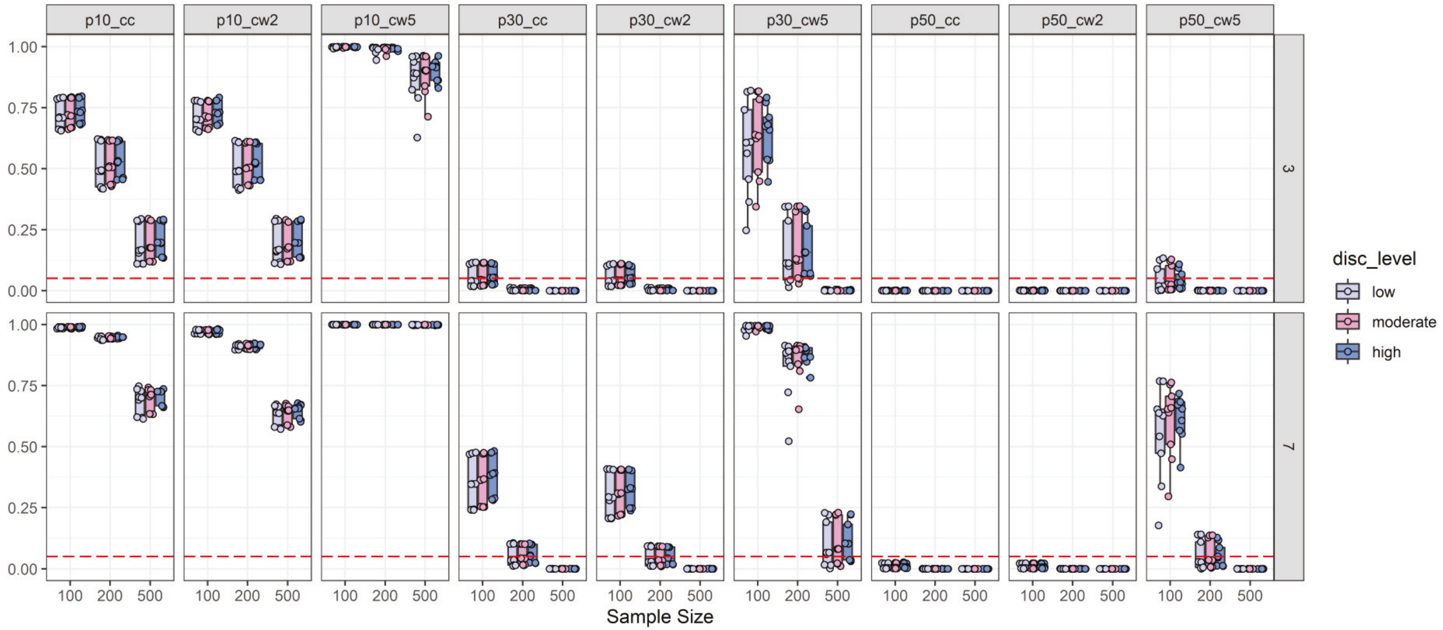


Figure A1. 2PL-discrimination split of type-I-error analysis: Depiction of the type-I-error rate (y-axis) as a function of sample size (x-axis), number of distractors (three distractors = top-row; seven distractors = bottom-row), 2PL-discrimination (disc_level: Low vs. moderate vs. high), and effect size measures combined with effect size thresholds. The horizontal red dashed line represents the target type-I-error rate of 0.05. For more explanations, see Figure 2.

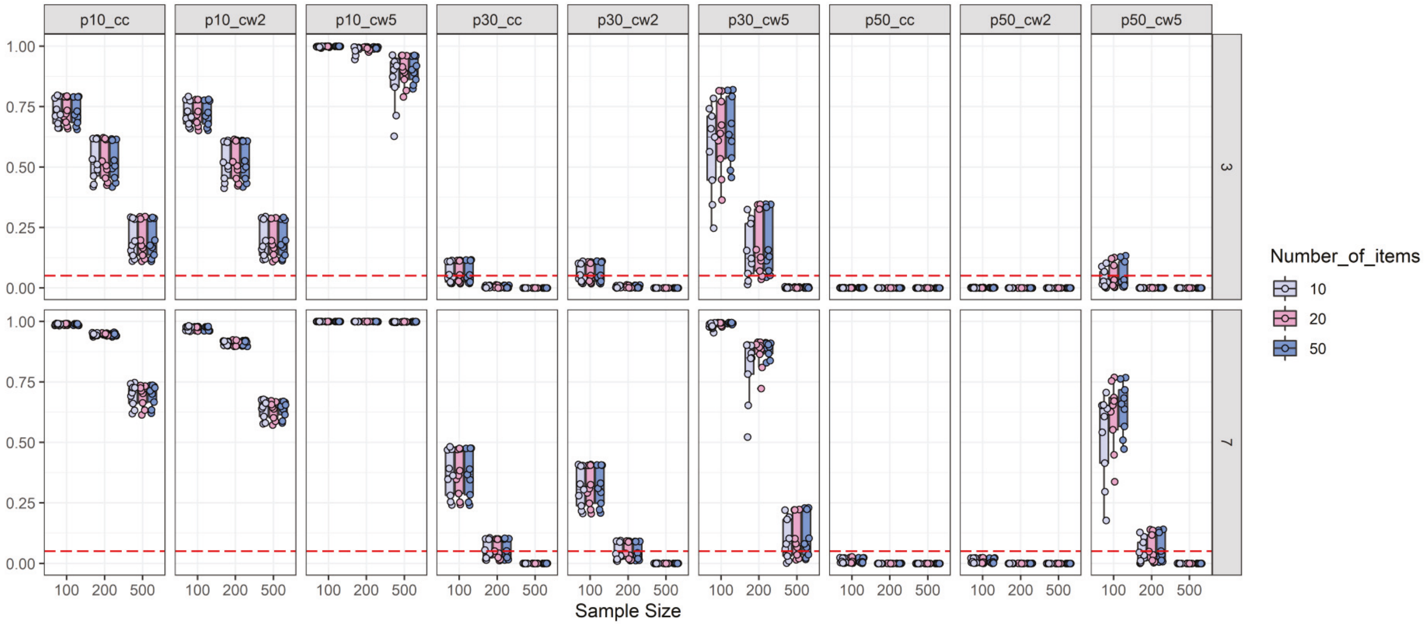


Figure A2. Number-of-items split of type-I-error analysis: Depiction of the type-I-error rate (y-axis) as a function of sample size (x-axis), number of distractors (three distractors = top-row; seven distractors = bottom-row), number of items and effect size measures combined with effect size thresholds. The horizontal red dashed line represents the target type-I-error rate of 0.05. For more explanations, see Figure 2.

Appendix B

In this appendix, further simulation results on the empirical power under boundary conditions are displayed. Figure A3 shows the findings for average γ as a boundary condition when the boxes in the boxplots are grouped by the levels of simulated 2PL discrimination. The power findings replicated well even with this boundary condition. Only low 2PL discrimination conditions were associated with clearly decreasing levels of power far below the target level of 0.80. The same observation was made with average PB_{DC} as a boundary condition (see Figure A4). 2PL discrimination was found to be the strongest influencing factor on these examinations of boundary conditions. For 2PL difficulty, a similar pattern was revealed with the lowest power for moderately difficult items (in some cases even dropping below the 0.80 target level) and the highest power for very difficult items (see Figures A5 and A6). When structuring the boxes in the boxplots according to the number of items, it was further revealed that the number of items was inversely related to the dispersion of power results (see Figures A7 and A8). With ten items power findings were pretty homogenous for all studied effect size measures, but for 50 items, the power results were strongly scattered.

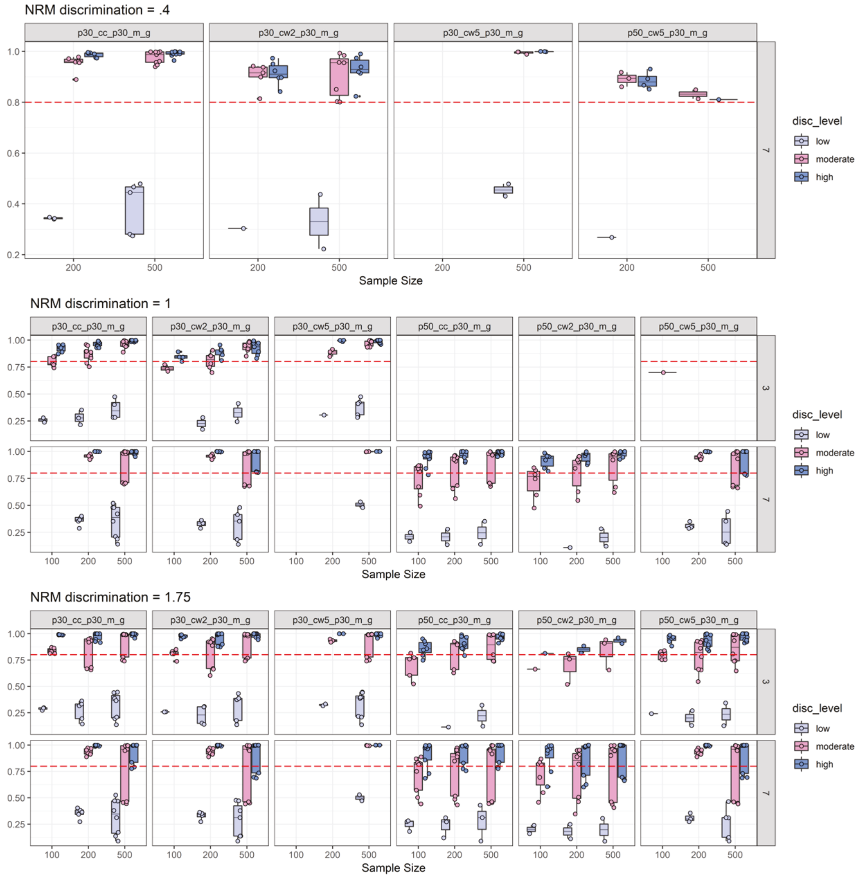


Figure A3. 2PL-discrimination split of empirical power analysis: Depiction of the empirical power (y-axis) under the boundary condition that the average γ is greater than 0.30. Power is depicted as a function of NRM discrimination, sample size (x-axis), number of distractors, 2PL-discrimination, and discrimination effect sizes combined with effect size thresholds (p30_cc_p30_m_g = canonical correlation with a 0.30 threshold; p30_cw2_p30_m_g = Cohen's ω based on two ability groups with a 0.30 threshold; p30_cw5_p30_m_g = Cohen's ω based on five ability groups with a 0.30 threshold; p50_cc_p30_m_g = canonical correlation with a 0.50 threshold; p50_cw2_p30_m_g = Cohen's ω based on two ability groups with a 0.50 threshold; p50_cw5_p30_m_g = Cohen's ω based on five ability groups with a 0.50 threshold). The horizontal red dashed line represents the target power level of 0.80. For more explanations, see Figures 2 and 3.

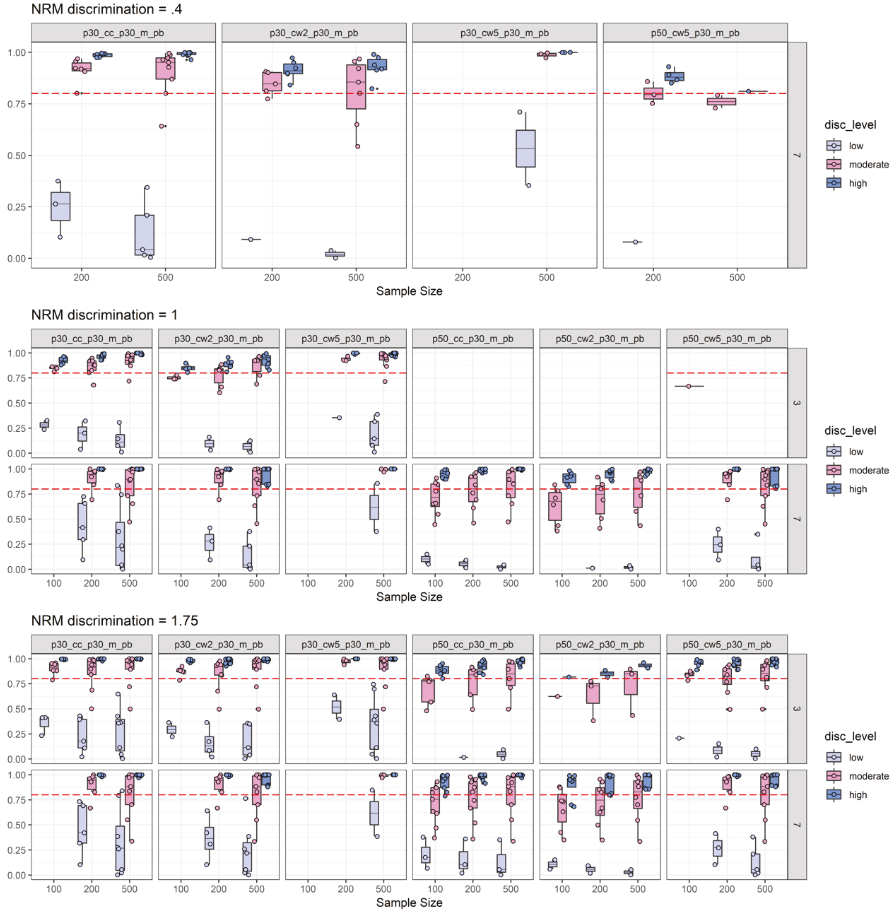


Figure A4. 2PL-discrimination split of empirical power analysis: Depiction of the empirical power (y-axis) under the boundary condition that the average PB_{DC} is smaller than −0.30. Power is depicted as a function of NRM discrimination, sample size (x-axis), number of distractors, 2PL-discrimination, and discrimination effect sizes combined with effect size thresholds (p30_cc_p30_m_pb = canonical correlation with a 0.30 threshold; p30_cw2_ p30_m_pb = Cohen's ω based on two ability groups with a 0.30 threshold; p30_cw5_ p30_m_pb = Cohen's ω based on five ability groups with a 0.30 threshold; p50_cc_ p30_m_pb = canonical correlation with a 0.50 threshold; p50_cw2_ p30_m_pb = Cohen's ω based on two ability groups with a 0.50 threshold; p50_cw5_ p30_m_pb = Cohen's ω based on five ability groups with a 0.50 threshold). The horizontal red dashed line represents the target power level of 0.80. For more explanations, see Figures 2 and 3.

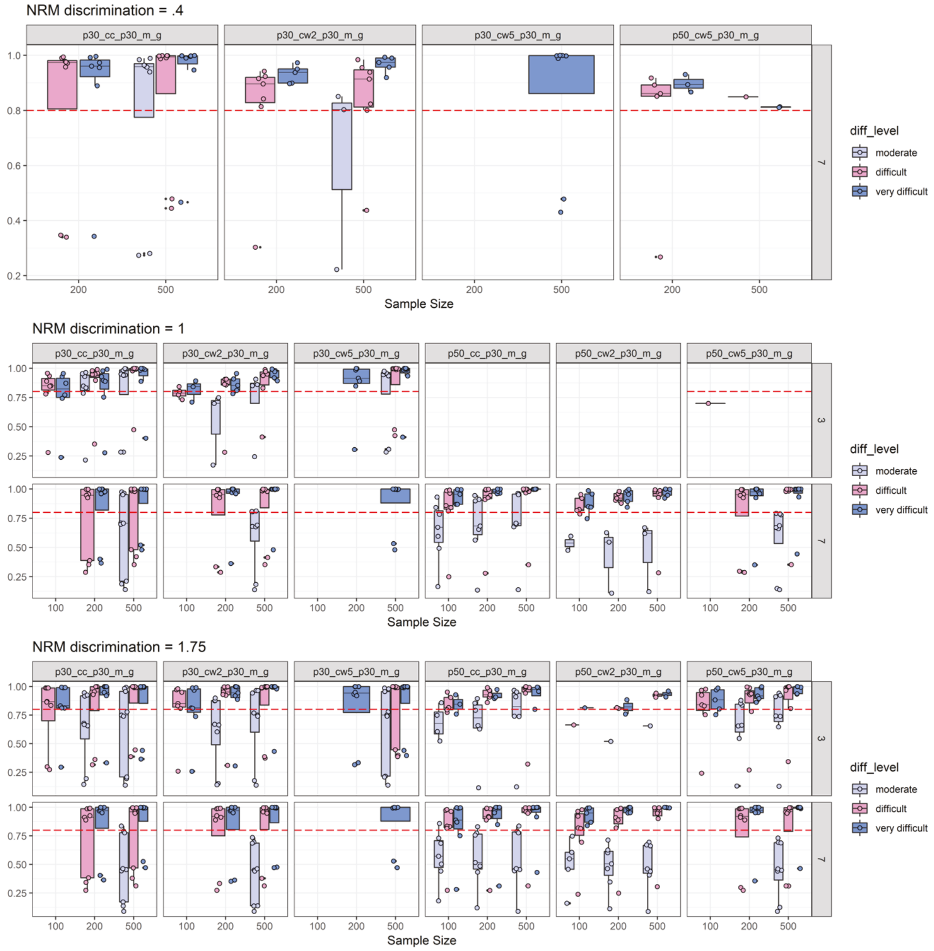


Figure A5. 2PL-difficulty split of empirical power analysis: Depiction of the empirical power (y-axis) under the boundary condition that the average γ is greater than 0.30. Power is depicted as a function of NRM discrimination, sample size (x-axis), number of distractors, 2PL-difficulty, and discrimination effect sizes combined with effect size thresholds. The horizontal red dashed line represents the target power level of 0.80. For more explanations, see Figures 2 and 3.

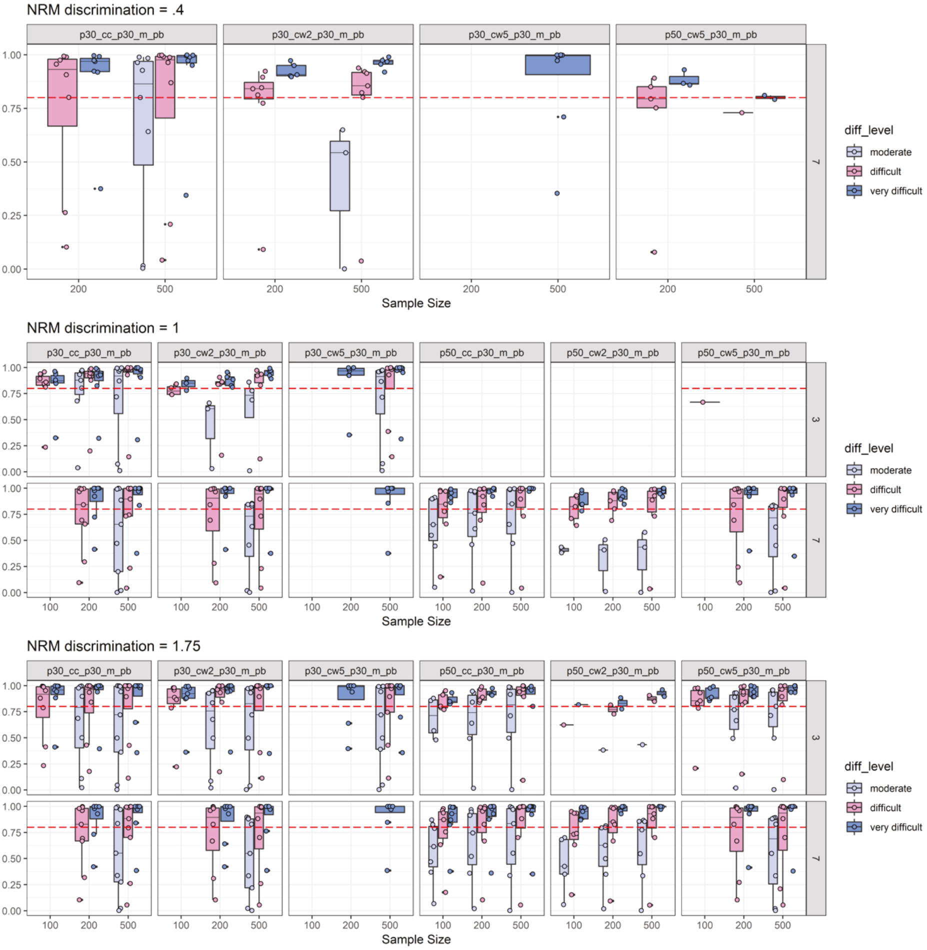


Figure A6. 2PL-difficulty split of empirical power analysis: Depiction of the empirical power (y-axis) under the boundary condition that the average PB_{DC} is smaller than −0.30. Power is depicted as a function of NRM discrimination, sample size (x-axis), number of distractors, 2PL-difficulty, and discrimination effect sizes combined with effect size thresholds. The horizontal red dashed line represents the target power level of 0.80. For more explanations, see Figures 2 and 3.

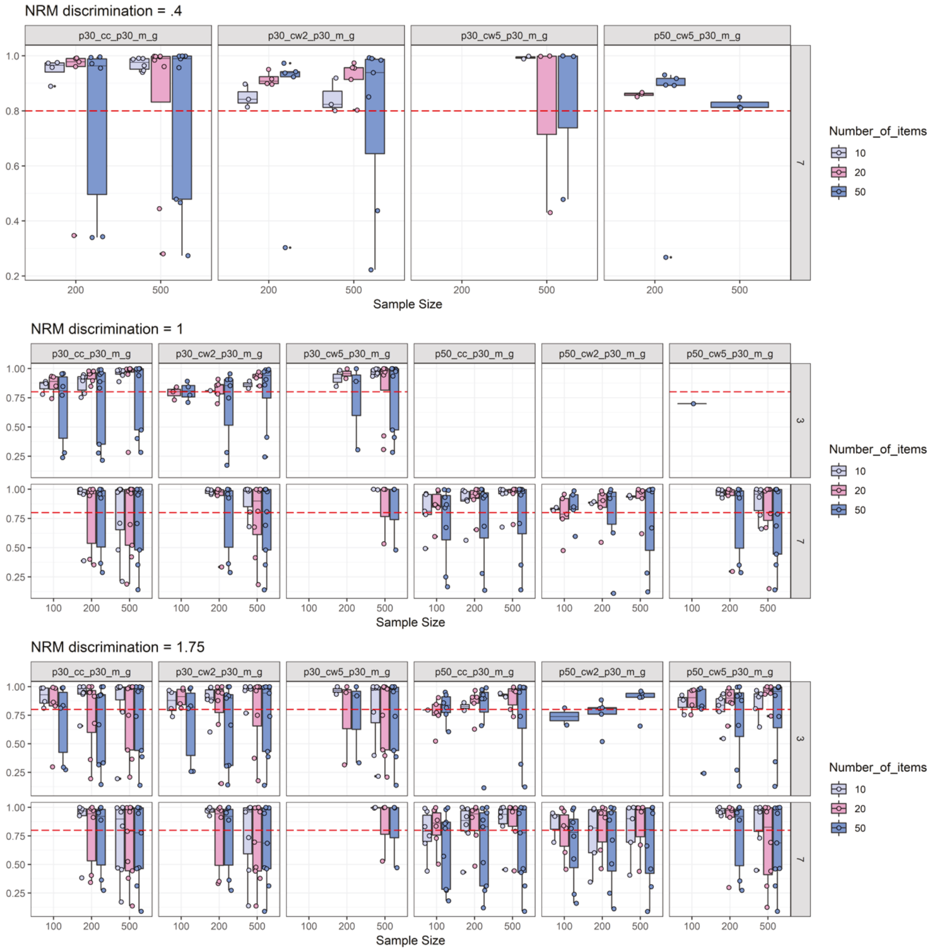


Figure A7. Number-of-items split of empirical power analysis: Depiction of the empirical power (y-axis) under the boundary condition that the average γ is greater than 0.30. Power is depicted as a function of NRM discrimination, sample size (x-axis), number of distractors, number of items, and discrimination effect sizes combined with effect size thresholds. The horizontal red dashed line represents the target power level of 0.80. For more explanations, see Figures 2 and 3.

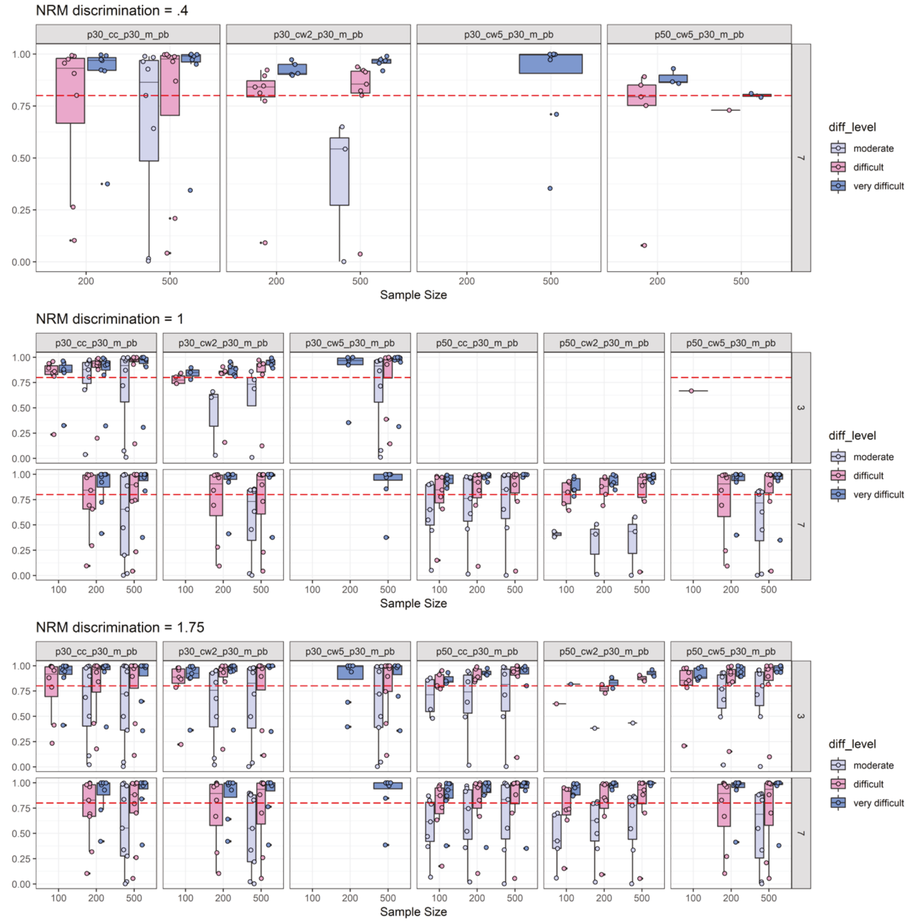


Figure A8. Number-of-items split of empirical power analysis: Depiction of the empirical power (y-axis) under the boundary condition that the average PB_{DC} is smaller than -0.30. Power is depicted as a function of NRM discrimination, sample size (x-axis), number of distractors, and discrimination effect sizes combined with effect size thresholds. The horizontal red dashed line represents the target power level of 0.80. For more explanations, see Figures 2 and 3.

Appendix C

In this appendix, the R code to reproduce the analysis for the Myszkowski-Storme dataset is presented. Two additional packages were used: Psych (Revelle 2018) for the scoring of the multiple-choice items, and mirt (Chalmers 2012) for estimation of the 2PL parameters. The complete R code, including also the simulation study is available in the OSF repository: https://osf.io/9tp8h/.

```
### load data
#
### downloaded from
# https://data.mendeley.com/datasets/h3yhs5gy3w/1
dataset <- read.csv(“dataset.csv”, stringsAsFactors=FALSE)
### install required packages (if needed)
### remove # to make this code run
#install.packages(c(“psych”,”mirt”))
### get results function
### includes also the effect size measures that were
### studied in the simulation and also more useful
### descriptive statistics
get_results <- function(data,keys=“sim”){
  ### keys for scoring
  if(length(keys)==1){keys <- rep(0,ncol(data))}else{
    keys <- keys
  }
  ### quantiles for two ability groups
  p2 <- .5
  ### quantiles for five ability groups
  p5 <- c(.2,.4,.6,.8)
  ### load psych library
  require(psych)
  ### score all items
  scored <- score.multiple.choice(key=keys,data=data,score=F)
  ### ability groups
  abil2.c <- rep(0,nrow(scored))
  for(i in 1:length(p2)){
  if(i < length(p2)){
    abil2.c[rowSums(scored)>quantile(rowSums(scored),p=p2[i])
            & rowSums(scored)<=quantile(rowSums(scored),p=p2[i+1])] <- i
  }else{abil2.c[rowSums(scored)>quantile(rowSums(scored),p=p2[i])] <- i
  }
}
### ability groups
abil5.c <- rep(0,nrow(scored))
for(i in 1:length(p5)){
  if(i < length(p5)){
    abil5.c[rowSums(scored)>quantile(rowSums(scored),p=p5[i])
            & rowSums(scored)<=quantile(rowSums(scored),p=p5[i+1])] <- i
  }else{abil5.c[rowSums(scored)>quantile(rowSums(scored),p=p5[i])] <- i
  }
}
### list distractors with relative frequency < .05
rf05 <- list()
for(j in 1:ncol(data)){
  rf05[[j]] <- table(data[,j][data[,j]!=keys[j]])/length(data[,j])<.05
}
### general Cohen’s w, 2 ability groups
chi_g2 <- list()
cw_g2 <- list()
tab_c2_l <- list()
zero_columns2 <- list()
for(k in 1:ncol(data)){
  tab_c2 <- matrix(table(data[,k][data[,k]!=keys[k]],abil2.c[data[,k]!=keys[k]])
   [!rf05[[k]],],ncol=length(unique(abil2.c[data[,k]!=keys[k]])))
  zero_columns2[[k]] <- colSums(tab_c2)==0
  tab_c2 <- tab_c2[,colSums(tab_c2)>0]
  tab_c2_l[[k]]<-tab_c2
  if(sum(!rf05[[k]])>=2){chi_g2[[k]] <- chisq.test(tab_c2)}else{
    chi_g2[[k]] <- NA
  }
  ### Cohen's w - general
  if(sum(!rf05[[k]])>=2){cw_g2[[k]] <- sqrt(sum(((chi_g2[[k]]$observed/sum(tab_c2)
  -chi_g2[[k]]$expected/sum(tab_c2))^2)/(chi_g2[[k]]$expected/sum(tab_c2))))}else{
    cw_g2[[k]] <- NA
  }
}
### general Cohen's w, 5 ability groups
chi_g5 <- list()
cw_g5 <- list()
tab_c5_l <- list()
zero_columns5 <- list()
for(k in 1:ncol(data)){
  tab_c5 <- matrix(table(data[,k][data[,k]!=keys[k]],abil5.c[data[,k]!=keys[k]])
  [!rf05[[k]],],ncol=length(unique(abil5.c[data[,k]!=keys[k]])))
  zero_columns5[[k]] <- colSums(tab_c5)==0
  tab_c5 <- tab_c5[,colSums(tab_c5)>0]
  tab_c5_l[[k]]<-tab_c5
  if(sum(!rf05[[k]])>=2){chi_g5[[k]] <- chisq.test(tab_c5)}else{
    chi_g5[[k]] <- NA
  }
  ### Cohen's w - general
  if(sum(!rf05[[k]])>=2){cw_g5[[k]] <- sqrt(sum(((chi_g5[[k]]$observed/sum(tab_c5)
  -chi_g5[[k]]$expected/sum(tab_c5))^2)/(chi_g5[[k]]$expected/sum(tab_c5))))}else{
    cw_g5[[k]] <- NA
  }
}
### canonical correlation
can_cor <- list()
ncol_mmat <- list()
for(k in 1:ncol(data)){
  ncol_mmat[[k]] <- if(sum(!rf05[[k]])>=2)
  {ncol(model.matrix(rowSums(scored[scored[,k]==0,-1])~-1+factor(data[,k]
  [scored[,k]==0]))[,!rf05[[k]]])}else{
    NA
  }
  can_cor[[k]] <- if(sum(!rf05[[k]])>=2)
  {cancor(rowSums(scored[scored[,k]==0,-k]),model.matrix(rowSums(scored[scored[,k]
  ==0,-1])~-1+factor(data[,k][scored[,k]==0]))[,!rf05[[k]]])$cor}else{
    NA
  }
}
### point-biserial coefficient PB_DC
pb_dc <- list()
### Goodman-Kruskal gamma
```

```
gkg <- list()
gkg_tab <- list()
### start loop
for(v in 1:ncol(data)){
  pb_dc_d <- list()
  gkg_d <- list()
  gkg_tab_d <- list()
  ### function to calculate
  ### Goodman-Kruskal gamma
  ### taken from here:
  ### https://stat.ethz.ch/pipermail/r-help/2003-March/030835.html
  goodman <- function(x,y){
    Rx <- outer(x,x,function(u,v) sign(u-v))
    Ry <- outer(y,y,function(u,v) sign(u-v))
    S1 <- Rx*Ry
    return(sum(S1)/sum(abs(S1)))}
  ### start loop
  non_key <- unique(data[,v])[!unique(data[,v])%in%keys[v]]
  for(w in non_key){
    MDC <- mean(rowSums(scored)[data[,v]%in%c(keys[v],w)])
    SDC <- sd(rowSums(scored)[data[,v]%in%c(keys[v],w)])
    MD <- mean(rowSums(scored)[data[,v]%in%w])
    PD <- mean(data[,v]%in%w)
    PC <- mean(data[,v]%in%keys[v])
    ### r-PB_D
    ### r-PB_DC
    pb_dc_d[[w]] <- (MD-MDC)/SDC*sqrt(PD/PC)
    ### Goodman-Kruskal gamma
    score_other_items <- factor(rowSums(scored[,-v]))
    tab_gkg_d <- table(data[data[,v]%in%c(keys[v],w),v],score_other_items[data[,v]
    %in%c(keys[v],w)])
    ### exclude ability levels with zero frequency
    tab_gkg_d <- tab_gkg_d[,colSums(tab_gkg_d)>0]
    gkg_d[[w]] <- goodman(as.numeric(colnames(tab_gkg_d)),
                        tab_gkg_d[as.numeric(rownames(tab_gkg_d))%in%keys[v],]
                        /colSums(tab_gkg_d))
    gkg_tab_d[[w]] <- tab_gkg_d
  }
  pb_dc[[v]] <- pb_dc_d
  gkg[[v]] <- gkg_d
  gkg_tab[[v]] <- gkg_tab_d
}
### return results
res <- list(rf05 = rf05,
            tab_c2_l = tab_c2_l, zero_columns2 = zero_columns2,
            tab_c5_l = tab_c5_l, zero_columns5 = zero_columns5,
            cw_g2 = cw_g2, cw_g5 = cw_g5,
            can_cor = can_cor,
            pb_dc = pb_dc,
            gkg = gkg,
            gkg_tab = gkg_tab,
            ncol_mmat = ncol_mmat)
return(res)
}
### frequencies of distractor usage
### including correct response
apply(dataset,2,table)
### load psych library
library(psych)
### score all items
scored <- score.multiple.choice(key=c(7,6,8,2,1,5,1,6,3,2,4,5),data=dataset,score=F)
### does choosing a certain other distractor
### imply better overall scores?
#
### run suggested distractor analysis
ms_res<-get_results(dataset,keys = c(7,6,8,2,1,5,1,6,3,2,4,5))
### show Results for Table 4
#
### show for which items the distractor choice frequencies were
### below 5%:
ms_res$rf05
### Items 1 to 5 have too many distractors with response frequencies
### below 5%.
#
### get 2PL parameters from mirt
library(mirt)
est_test2pl <- mirt(scored, 1, itemtype="2PL")
### show results
coef(est_test2pl)
# a1 = 2PL-discrimination
# d  = 2PL-difficulty
#
### canonical correlation findings
ms_res$can_cor[6:12]
### check boundary conditions
#
### average pb_dc
lapply(ms_res$pb_dc,function(x)mean(unlist(x),na.rm=T))[6:12]
### average gamma
lapply(ms_res$gkg,function(x)mean(unlist(x),na.rm=T))[6:12]
```

References

Arendasy, Martin, and Markus Sommer. 2013. Reducing response elimination strategies enhances the construct validity of figural matrices. *Intelligence* 41: 234–43. [CrossRef]

Arendasy, Martin, Markus Sommer, Georg Gittler, and Andreas Hergovich. 2006. Automatic generation of quantitative reasoning items: A pilot study. *Journal of Individual Differences* 27: 2–14. [CrossRef]

Attali, Yigal, and Tamar Fraenkel. 2000. The point-biserial as a discrimination index for distractors in multiple-choice items: Deficiencies in usage and an alternative. *Journal of Educational Measurement* 37: 77–86. [CrossRef]

Barton, Mark A., and Frederic M. Lord. 1981. An upper asymptote for the three-parameter logistic item-response model. *ETS Research Report Series* 1: i-8. [CrossRef]

Bethell-Fox, Charles E., David F. Lohman, and Richard E. Snow. 1984. Adaptive reasoning: Componential and eye movement analysis of geometric analogy performance. *Intelligence* 8: 205–38. [CrossRef]

Blum, Diego, and Heinz Holling. 2018. Automatic generation of figural analogies with the IMak package. *Frontiers in Psychology* 9: 1286. [CrossRef]

Blum, Diego, Heinz Holling, Maria S. Galibert, and Boris Forthmann. 2016. Task difficulty prediction of figural analogies. *Intelligence* 56: 72–81. [CrossRef]

Bock, R. Darrell. 1972. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika* 37: 29–51. [CrossRef]

Chalmers, R. Philip. 2012. mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software* 48: 1–29. [CrossRef]

Cohen, J. 1992. A power primer. *Psychological Bulletin* 112: 155–59. [CrossRef]

Crocker, Linda S., and James Algina. 1986. *Introduction to Classical and Modern Test Theory*. Forth Worth: Harcourt Brace Jovanovich.

Cureton, Edward. 1966. Corrected item-test correlations. *Psychometrika* 31: 93–96. [CrossRef]

Davis, Frederick B., and Gordon Fifer. 1959. The effect on test reliability and validity of scoring aptitude and achievement tests with weights for every choice. *Educational and Psychological Measurement* 19: 159–70. [CrossRef]

DeMars, Christine E. 2003. Sample size and recovery of nominal response model item parameters. *Applied Psychological Measurement* 27: 275–88. [CrossRef]

Garcia-Perez, Miguel A. 2014. Multiple-choice tests: Polytomous IRT models misestimate item information. *Spanish Journal of Psychology* 17: e88. [CrossRef] [PubMed]

Gierl, Mark J., Okan Bulut, Qi Guo, and Xinxin Zhang. 2017. Developing, analyzing, and using distractors for multiple-choice tests in education: A comprehensive review. *Review of Educational Research* 87: 1082–116. [CrossRef]

Gonthier, Corentin, and Jean-Luc Roulin. 2019. Intraindividual strategy shifts in Raven's matrices, and their dependence on working memory capacity and need for cognition. *Journal of Experimental Psychology: General* 149: 564–79. [CrossRef] [PubMed]

Gonthier, Corentin, and Noémylle Thomassin. 2015. Strategy use fully mediates the relationship between working memory capacity and performance on Raven's matrices. *Journal of Experimental Psychology: General* 144: 916–24. [CrossRef] [PubMed]

Goodman, Leo A., and William H. Kruskal. 1979. *Measures of Association for Cross Classifications*. New York: Springer.

Guttman, Louis, and Izchak M. Schlesinger. 1967. Systematic construction of distractors for ability and achievement test items. *Educational and Psychological Measurement* 27: 569–80. [CrossRef]

Haladyna, Thomas M. 2004. *Developing and Validating Multiple-Choice Test Items*. New York: Routledge.

Haladyna, Thomas M., and Steven M. Downing. 1993. How many options is enough for a multiple-choice test item? *Educational and Psychological Measurement* 53: 999–1010. [CrossRef]

Harville, David A. 2008. *Matrix Algebra from a Statistician's Perspective*. New York: Springer.

Hayes, Taylor R., Alexander A. Petrov, and Per B. Sederberg. 2011. A novel method for analyzing sequential eye movements reveals strategic influence on Raven's Advanced Progressive Matrices. *Journal of Vision* 11: 1–11. [CrossRef]

Henrysson, Sten. 1962. The relation between factor loadings and biserial correlations in item analysis. *Psychometrika* 27: 419–29. [CrossRef]

Henrysson, Sten. 1971. Gathering, analyzing, and using data on test items. In *Educational Measurement*, 2nd ed.; Edited by Robert L. Thorndike. Beverly Hills: American Council on Education.

Hornke, Lutz F., and Michael W. Habon. 1986. Rule-based item bank construction and evaluation within the linear logistic framework. *Applied Psychological Measurement* 10: 369–80. [CrossRef]

Jacobs, Paul I., and Mary Vandeventer. 1970. Information in wrong responses. *Psychological Reports* 26: 311–15. [CrossRef]

Jarosz, Andrew F., and Jennifer Wiley. 2012. Why does working memory capacity predict RAPM performance? A possible role of distraction. *Intelligence* 40: 427–38. [CrossRef]

Johanson, George A., and Gordon P. Brooks. 2010. Initial scale development: Sample size for pilot studies. *Educational Psychological Measurement* 70: 394–400. [CrossRef]

Klecka, William R. 1980. *Discriminant Analysis*. Beverly Hills: SAGE Publications, ISBN 0-8039-1491-1.

Kline, Paul. 2000. *The Handbook of Psychological Testing*. London: Routledge.

Kunda, Maithilee, Isabelle Soulieres, Agata Rozga, and Ashok K. Goel. 2016. Error patterns on the Raven's Standard Progressive Matrices test. *Intelligence* 59: 181–98. [CrossRef]

Levine, Michael V., and Fritz Drasgow. 1983. The relation between incorrect option choice and estimated ability. *Educational Psychological Measurement* 43: 675–85. [CrossRef]

Lord, Frederic M. 1980. *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale: Lawrence-Erlbaum Associates.

Love, Thomas E. 1997. Distractor selection ratios. *Psychometrika* 62: 51–62. [CrossRef]

Matzen, Laura E., Zachary O. Benz, Kevin R. Dixon, Jamie Posey, James K. Kroger, and Ann E. Speed. 2010. Recreating Raven's: Software for systematically generating large numbers of Raven-like matrix problems with normed properties. *Behavor Research Methods* 42: 525–41. [CrossRef]

Mitchum, Ainsley L., and Colleen M. Kelley. 2010. Solve the problem first: Constructive solution strategies can influence the accuracy of retrospective confidence judgments. *Journal of Experimental Psychology: Learning, Memory, Cognition* 36: 699–710. [CrossRef]

Muraki, Eiji. 1992. A generalized partial credit model: Application of an EM algorithm. *ETS Research Report Series* 1: i-30. [CrossRef]

Myszkowski, Nils, and Martin Storme. 2018. A snapshot of g? Binary and polytomous item-response theory investigations of the last series of the Standard Progressive Matrices (SPM-LS). *Intelligence* 68: 109–16. [CrossRef]

Nunnally, Jum C., and Ira H. Bernstein. 1994. *Psychometric Theory*. New York: McGraw-Hill.

R Core Team. 2019. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Revelle, William. 2018. *Psych: Procedures for Personality and Psychological Research*. Evanston: Northwestern University.

Revuelta, Javier. 2005. An item response model for nominal data based on the rising selection ratios criterion. *Psychometrika* 70: 305–24. [CrossRef]

Schiano, Diane J., Lynn A. Cooper, Robert Glaser, and Hou C. Zhang. 1989. Highs are to lows as experts are to novices: Individual differences in the representation and solution of standardized figural analogies. *Human Performance* 2: 225–48. [CrossRef]

Sigel, Irving E. 1963. How intelligence tests limit understanding of intelligence. *Merrill-Palmer Quarterly of Behavior and Development* 9: 39–56.

Snow, Richard E. 1980. Aptitude processes. In *Aptitude, Learning, and Instruction: Cognitive Process Analyses of Aptitude*. Edited by Richard E. Snow, Pat-Anthony Federico and William E. Montague. Hillsdale: Erlbaum, vol. 1, pp. 27–63. ISBN 978-089-859-043-2.

Storme, Martin, Nils Myszkowski, Simon Baron, and David Bernard. 2019. Same test, better scores: Boosting the reliability of short online intelligence recruitment tests with nested logit item response theory models. *Journal of Intelligence* 7: 17. [CrossRef]

Suh, Youngsuk, and Daniel M. Bolt. 2010. Nested logit models for multiple-choice item response data. *Psychometrika* 75: 454–73. [CrossRef]

Thissen, David. 1976. Information in wrong responses to the Raven Progressive Matrices. *Journal of Educational Measurement* 13: 201–14. [CrossRef]

Thissen, David, Lynne Steinberg, and Anne R. Fitzpatrick. 1989. Multiple-choice models: The distractors are also part of the item. *Journal of Educational Measurement* 26: 161–76. [CrossRef]

Thompson, Bruce. 1984. *Canonical Correlation Analysis*. Newbury Park: SAGE Publications, ISBN 0-8039-2392-9.

Vejleskov, Hans. 1968. An analysis of Raven matrix responses in fifth grade children. *Scandinavian Journal Psychology* 9: 177–86. [CrossRef]

Vigneau, François, André F. Caissie, and Douglas A. Bors. 2006. Eye-movement analysis demonstrates strategic influences on intelligence. *Intelligence* 34: 261–72. [CrossRef]

Vodegel Matzen, Linda B. L., Maurits W. van der Molen, and Ad C. M. Dudink. 1994. Error analysis of Raven test performance. *Personality and Individual Differences* 16: 433–45. [CrossRef]

Von der Embse, Nathaniel P., Andrea D. Mata, Natasha Segool, and Emma-Catherine Scott. 2014. Latent profile analyses of test anxiety: A pilot study. *Journal of Psychoeducational Assessessment* 32: 165–72. [CrossRef]

Wainer, Howard. 1989. The future of item analysis. *Journal of Educational Measurement* 26: 191–208. [CrossRef]

Yen, Wendy M., and Anne R. Fitzpatrick. 2006. Item Response Theory. In *Educational Measurement*. Edited by Robert L. Brennan. Westport: Praeger Publishers.

Journal of
Intelligence

MDPI

Article

Diagnosing a 12-Item Dataset of Raven Matrices: With Dexter

Ivailo Partchev

Cito, 6814 CM Arnhem, The Netherlands; Ivailo.Partchev@cito.nl

Received: 21 February 2020; Accepted: 27 April 2020; Published: 6 May 2020

check for updates

Abstract: We analyze a 12-item version of Raven's Standard Progressive Matrices test, traditionally scored with the sum score. We discuss some important differences between assessment in practice and psychometric modelling. We demonstrate some advanced diagnostic tools in the freely available R package, dexter. We find that the first item in the test functions badly—at a guess, because the subjects were not given exercise items before the live test.

Keywords: intelligence tests; classical test theory; IRT; interaction model; test-item regression

1. Introduction

Myszkowski and Storme (2018) have applied a number of binary and polytomous item-response theory (IRT) Lord (1980) models to the last series of Raven's Standard Progressive Matrices (SPM) test Raven (1941), further referred to as the SPM-LS test. They have made their dataset publicly available, and the *Journal of Intelligence* has proposed a special issue where other researchers are encouraged to present their own analyses.

The idea is not entirely new. Back in 1976, Thissen (1976) tried to apply Bock's nominal response model Bock (1972) to Raven's matrices as an attempt to throw light on the functioning of the distractors and improve scoring in the lower ability range. It is easy to overlook this publication as it came so incredibly early, some five years before Bock and Aitkin (1981) proposed a really practicable way to estimate the model.

To start with the big question of whether applying complex IRT models to an old, venerable test of intelligence should be an improvement: I have not one but two answers. One is "possibly", the other "certainly not". The duplicity arises from the fact that it is not possible to have methods and criteria that would be equally appropriate to summative assessment, formative assessment, survey research, methodological research, or substantive research.

Consider assessment. Computer-assisted learning has developed at staggering rates, becoming essentially intertwined with formative assessment. Operating within the effort to increase ability, we can even enjoy the luxury of being able to ask the same item multiple times and observe learning happen. Summative assessment has remained more traditional: We tend to interrupt the learning process for a while, hoping that ability will remain unchanged during testing, and praying that the items have not been compromised by disclosure. The two modes are not simply different—they are more like opposites. Hence, there is no methodological one-size-fits-all—not even within assessment practice.

On the other hand, not everybody who analyzes test data is busy grading exams. Some might be studying populations, as is the case with PISA, TIMSS and friends. Others might be interested in the way people behave when answering educational or intelligence tests. They will come up with ideas and hypotheses whose evidential support will have to be demonstrated, since statements are not limited to a specific individual or projected to a specific finite population but generalized beyond. Goodness of fit plays a very different role in such circumstances than in the more artisanal job of making a measurement instrument for testing.

In the role of researchers, we might for example ask whether persons are guessing responses at random, and we can try to formalize the question into a testable model. It is a perfectly valid discussion Azevedo (2009); Glas (2009); Maris and Bechger (2009); Partchev (2009); San Martín et al. (2009); Thissen (2009); von Davier (2009) whether such a model, say the 3PL, is a good idea from a substantive or mathematical point of view. From my participation in that dispute it is clear that I am not very enthusiastic; see also Appendix B for some results in applying the 3PL model on the SPM-LS dataset. However, this is not the same as porting the 3PL model into assessment practice, the latter being predominantly ruled by the sum score. This is mainly for two reasons: (i) The sum score makes sense in a particular social situation and (ii) it seems to capture most of the essential information in the responses.

As Dorans (2012) notes, commenting on earlier work by Paul Holland, test takers can assume multiple roles: Those of learners, examinees, or contestants. Quoting from his abstract: “Test takers who are contestants in high-stakes settings want reliable outcomes obtained via acceptable scoring of tests administered under clear rules.” Telescoping to sports, where fairness is also a major issue, the 2020 edition of the ATP rulebook ATP Tour Inc. (2020) defines every conceivable rule and situation in the game of tennis on 374 pages (beats the APA Publication Manual American Psychological Association (2010) by more than 100 pages). Nothing is left to chance, everything is specified well before the game starts, and just how bizarre the idea that the scoring rules might be defined post hoc, based on a fairly opaque analysis of the results, and placidly assuming that athletes cheat as a rule. However, this is exactly what the 3PL model proposes.

Similar objections may be raised against the idea to ’exploit’ the potentially useful information in the wrong responses by fitting a nominal response model. Investigate in research—yes; exploit in assessment—rather not. When we are to pass judgement over individuals, our thinking tends to be more binary: Either the distractors are wrong and should get no credit, or they are sensible and should get partial credit. In either case, it should be part of the rules before the referee shouts “Time!”.

The need for simple scoring rules that are known before testing has begun, are easily explained to all parties involved, and are widely accepted as fair, is one of the main reasons why most assessment programs tend to rely on the sum score. When the test has more than one form, the choice is mainly between classical test theory (CTT) and equipercentile or kernel equating (still a hot topic, to judge by the number of recent books González and Wiberg 2017; Kolen and Brennan 2014; von Davier 2011; von Davier et al. 2004), or IRT, which provides an alternative solution to the equating problem. However, we would be interested primarily in models with sufficient statistics, such as the Rasch or the partial credit model, because they preserve the scoring rule (in the case of one test form, the ability estimates are just a monotone transformation of the sum score).

Another important advantage is that the degree of misfit of the IRT model would indicate the extent to which our scoring rule misses out potentially useful information. This is more realistic on the item level, where it can be a valuable tool in quality assurance. At test level and within IRT, it is more difficult to demonstrate misfit in practice (see also Appendix C). The search for that important thing that is not already captured by the sum score has become something of a Holy Grail in psychometrics—since the day when they added a second parameter to the Rasch model and up to the latest advances in cognitive diagnostic assessment Leighton and Gierl (2007). I have followed with great interest, have often been disappointed, and will probably be just as enthusiastic when the next wave appears.

What follows is an example of the initial data crunching that would be done at an educational testing institute when the data from a new test comes in. A careful exploratory analysis should always precede whatever comes next, whether assessment or further modelling and research; and we should not forget that the properties of an instrument and the properties of a dataset collected with it are not the same thing.

While playing Sherlock Holmes with the SPM-LS data, I take the opportunity to present our freely available R package, dexter, Maris et al. (2019) because it has been developed especially for this purpose and combines traditional and novel methods. The accent is on assessing and understanding item fit. There is no attempt at an exhaustive analysis of the psychometric properties of the 12-item test form, SPM-LS. Raven's matrices have been around for about 80 years and much is known about them—for example, Brouwers et al. (2009) examine 798 applications in 45 countries (N = 244,316) published between 1944 and 2003. Besides, an insight into the properties of the short form can be seen as the collective endeavour of the whole special issue—see, for example, Garcia-Garzon et al. (2019) for a factor-analytic analysis that shows the SPM-LS to be essentially unidimensional.

2. Materials and Methods

2.1. Data

The data is as supplied with the original study by Myszkowski and Storme (2018): The responses of 499 French undergraduate students aged between 19 and 24 to the twelve items of SPM-LS.

2.2. Methods

All analyses have been performed with dexter Maris et al. (2019), a freely available package for R Core Team (2013). Dexter has been created to be as useful as possible to both researchers and test practitioners, as long as they stay with models that have sufficient statistics for their parameters Andersen (1973). Every dexter project starts, as appropriate for testing, with a complete enumeration of the scoring rules for each item: Every admissible response gets mapped to an integer, with 0 as the lowest item score. Out of these rules, the program creates automatically a state-of-the-art relational data base optimized for the typical structure of test data.

The toolbox for assessing the quality of the items includes, among others:

- the usual statistics of classical test theory (CTT) Lord and Novick (1968);
- distractor plots, i.e., nonparametric regressions of each response alternative on the sum score;
- item-total regressions obtained directly from the data, from the calibration model (Rasch or partial credit), and from Haberman's interaction model Haberman (2007).

There is a companion package, dextergui Koops et al. (2019), providing an easy graphical user interface (GUI) to the basic functions. The GUI is very convenient: All tables are live, they can be sorted on each column, and clicking anywhere on the table opens up the appropriate graphs. However, in a paper like this it is easier to reproduce a script (see Appendix A) than to explain a GUI.

Readers of this journal will hardly need CTT statistics like item facility and item-total correlation, the Rasch model Rasch [1960] (1980), or the partial credit model (PCM) Masters (1982) explained. What we call distractor plots are non-parametric regressions of response alternatives on the total score. We produce them by estimating the density of the total score, overall and for each response alternative, and applying Bayes' rule to obtain the density of each response alternative given the total score.

A useful and novel method is a plot (example shown in Figure 1) that compares three item-total regressions:

- the empirical regression, shown with pink dots and representing, simply, the proportion of correct responses to the item (or the mean item score, for partial credit items), at each test score;
- the regression predicted by the Rasch (or partial credit) model, shown as a thin black line;

- the regression predicted by Haberman's interaction model, shown as a thicker gray line.

Item-total regressions (ITR) are somewhat similar to item response functions (IRF), but there are some important differences. The IRF combines an unobservable quantity on an arbitrary scale (on the x scale) with observable data (on the y axis) while the ITR only involves observable data.

What, however, is the interaction model? Well hidden in a book on an entirely different subject, Haberman's interaction model Haberman (2007) remains relatively unpopular and underestimated. We (the developers of dexter) have found it to be a very useful diagnostic tool, and we have generalized it to also handle polytomous items. The interaction model can be seen equivalently as a model for locally dependent items, a Rasch model where difficulty depends on item and score, and an exponential family model for classical test theory, as can be seen from the following equations:

$$P(\mathbf{x}|\theta) \propto \exp(\theta x_+ - \sum_i \beta_i x_i + \sum_i \sum_{j>i} (\sigma_i + \sigma_j) x_i x_j) \tag{1}$$

$$P(\mathbf{x}|\theta) \propto \exp(\theta x_+ - \sum_i (\beta_i + \sigma_i x_+) x_i) \tag{2}$$

$$P(\mathbf{X}|\theta) \propto \exp\left(\sum_i \beta_i x_{+i} + \sum_i \sigma_i \sum_p x_{pi} x_{p+} + \sum_s n_s \ln \lambda_s\right) \tag{3}$$

where i and j index items, p indexes persons, s indexes sum scores, and $+$ stands for summation. x are observed item responses, $\mathbf{x}$ a response vector, and $\mathbf{X}$ a matrix of responses. θ are latent abilities, β item difficulties, and σ are the item-specific interaction parameters featured in Haberman's model. The λ are there to reproduce (i.e., perfectly fit) the score distribution, and may be called score-parameters.

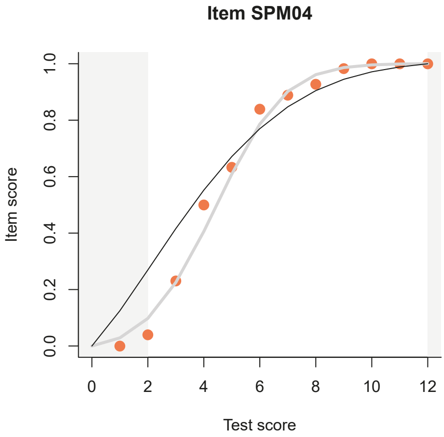


Figure 1. Example plot comparing three item-total regressions for the fourth item. Pink dots show the observed regression (in this case, proportion of correct responses at each distinct total score), predictions from the Rasch model are shown with a thin black line, and those from the interaction model with a thick gray line.

Each of these three representations can serve as the point of departure for a potentially useful discussion. Because our interest here is mainly in item fit, we will concentrate on the third one. We observe that the three terms in the exponential ensure that the model will reproduce perfectly, through the three sets of parameters, β, σ, and λ, the classical item facilities, the correlations of the item scores with the total score, and the distribution of the total scores. Note that this is more or less everything that we want to know about the data within CTT.

Let us return to Figure 1. I have deliberately chosen the item that deviates the most from the Rasch model in having a higher discrimination. The IM readily detects that, in fact, the 2PL model can be shown to be a low-rank approximation to the IM, so we have even more flexibility with the IM than with the 2PL model. However, unlike the two-, three- or many-PL models, the IM has sufficient statistics, it can be estimated via the conditional likelihood, and it makes predictions conditional on the observed total score, not on a hypothesized, latent quantity. This makes it much more appropriate for evaluating item fit.

Observe how, when the Rasch model and the IM deviate for an item, the pink dots representing the empirical item-total regression tend to cluster around the IM curve. This is what one typically sees in practice, and the points tend to get closer to the line as the sample size increases. In other words, not only does the IM reproduce exactly the three most interesting aspects of the response data from a CTT point of view, but it seems to capture all systematic deviations from the Rasch model, leaving out just the random noise. To make the plots even more legible, we have introduced 'curtains' slightly obscuring but not hiding the lower and upper 5% of the data as measured on the test score. This helps concentrate on the really important differences among the three ITR.

Neither the Rasch model nor the IM make any provisions for random guessing. The 3PL model, on the contrary, assumes that people always guess, and then tries to fit a curve with a particular shape to the data. Even if that is successful (the model has an identification problem, as shown in Azevedo 2009; Glas 2009; Maris and Bechger 2009; Partchev 2009; San Martín et al. 2009; Thissen 2009; von Davier 2009), the data can lie near to the curve for many reasons, one of which is random guessing. None of the three models have a device to tell us whether people are actually guessing or not.

The two smoothed ITR start and end at the same points as the observed ITR. Inevitably, both the observed and the predicted item score must be 0 when the total score is 0, and when a person achieves a full total score, the item score for each item must also take the maximum possible value. This gives a specific aspect to the ITR depending on the slope. When an item discriminates better than predicted by the Rasch model, the ITR of the IM retains the same sigmoid shape but gets steeper. When discrimination is low, typical of badly written items, the curve starts to bend, resembling a cubic polynomial. This is particularly expressive when the ITR must accommodate a negative slope in the middle, typical of items with a wrong answer key. When the slope is small or negative, the ITR of the IM suggests that persons of low ability (say, at the left curtain) have a spuriously high probability of a correct response. This is not necessarily due to guessing.

To summarize: I believe that items discriminating similar to or better than what is expected under the Rasch model can be used without consternation: Differences in the slope will cancel when we sum together even a very modest number of item scores (see also Appendix C). Low discrimination always means trouble of one kind or another. So, my recommended workflow is to catch such items, starting with the item-total and item-rest correlations and proceeding with the item-total regressions. A careful analysis of the distractor plots for the offending items will help diagnose what is wrong with the item and suggest a revision.

3. Results

The 12-item test, SPM-LS, has a decent Cronbach alpha of 0.81, and an average item facility of 0.65. The average correlation with the total score (rit) is 0.57, and the average correlation with the rest score (rir) is 0.47.

Table 1 shows the essential item level CTT statistics. As expected from the structure of the SPM test, the item facilities progressively decrease with the notable exception of the first item. Discrimination, as characterized by the rit and the rir, is highest in the middle and lowest at both ends, which is what one would expect from point-biserial correlations. However, the discrimination for the first item is a bit too low, especially as the item does not appear to be as easy as anticipated.

Table 1. Selected item level CTT statistics for the SPM-LS data set.

Item	Facility	rit	rir
SPM01	0.76	0.43	0.30
SPM02	0.91	0.48	0.40
SPM03	0.80	0.56	0.46
SPM04	0.82	0.66	0.58
SPM05	0.86	0.65	0.57
SPM06	0.76	0.66	0.56
SPM07	0.70	0.59	0.47
SPM08	0.58	0.63	0.52
SPM09	0.57	0.57	0.44
SPM10	0.39	0.63	0.51
SPM11	0.36	0.55	0.42
SPM12	0.32	0.48	0.34

These observations are facilitated by the plots on Figure 2.

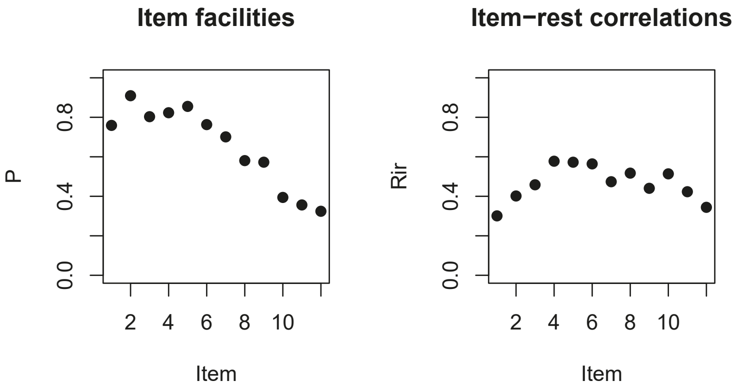


Figure 2. Item facility (**left**) and correlation with the rest score (**right**) by position of the item in the SPM-LS test.

The slight hint that the first item, SPM01, may be out of line with the others, becomes quite dramatic when we examine the ITR plots (Figure 3). Just compare the plots for the first two items, which are supposed to be very similar. For item SPL02, the Rasch model and the IM agree almost perfectly. According to both models, persons of low ability, say at the left curtain (fifth percentile) have a high probability, well over 0.5, to answer correctly, but this is obviously because the item is very easy.

The plot for item SPM01 is very different. The IM curve has a very particular and irregular shape, thanks to which it deviates sharply from the Rasch model in the lower ability range, but much less in the upper range. What is going on among persons of lower ability? Are they guessing? The pseudo-guessing parameter in the 3PL model (Appendix B) is equal to zero for both items, and what is the logic to guess when the item is so easy? Why on the first item but not on the second?

Figure 4 shows distractor plots (non-parametric regressions of each response alternative to an item on the total test score), which are a less sophisticated but very detailed and useful diagnostic tool when an item appears spurious on the traditional CTT statistics and/or ITR plots. I have included the plots for all 12 items because an anonymous reviewer suggested that the reader would like to see them, and I completely agree; however, the idea is that, in practice, the CTT statistics and the ITR plots will help us narrow down the detailed diagnostic plots that we need to examine to the most problematic items.

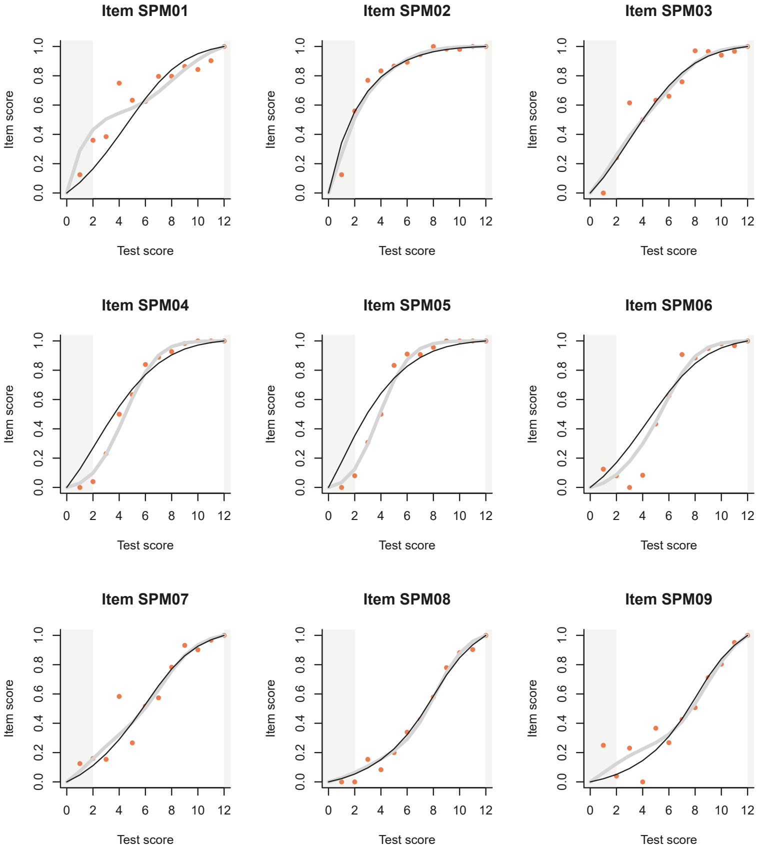


Figure 3. *Cont.*

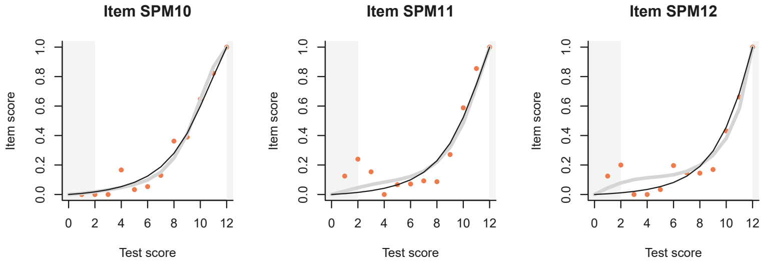


Figure 3. Item-total regressions for the items in the SPM-LS test obtained from the data (pink dots), the Rasch model (thin black lines), and the interaction model (thick gray lines).

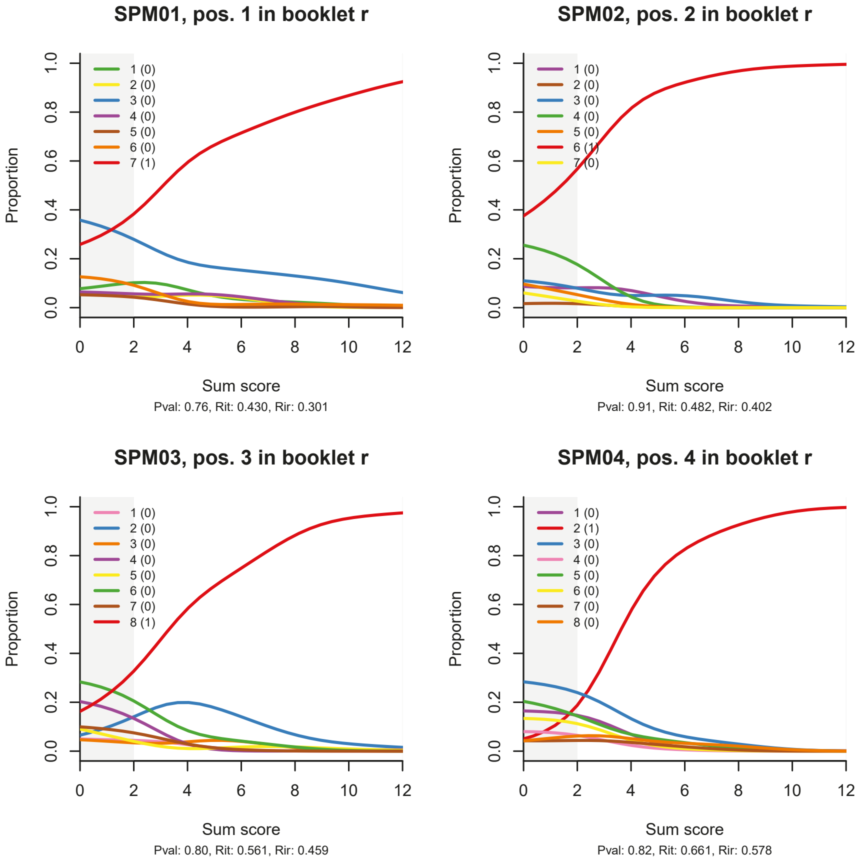


Figure 4. *Cont.*

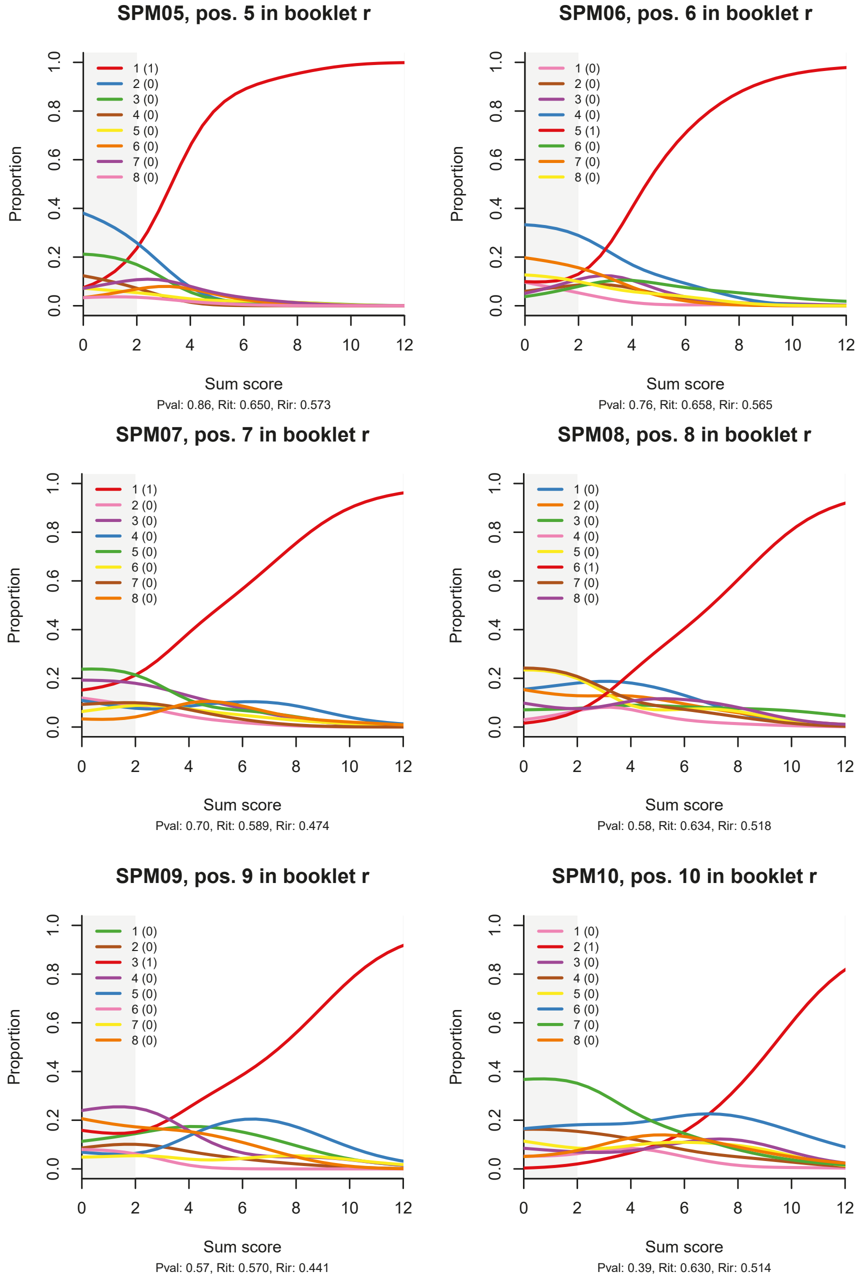


Figure 4. *Cont.*

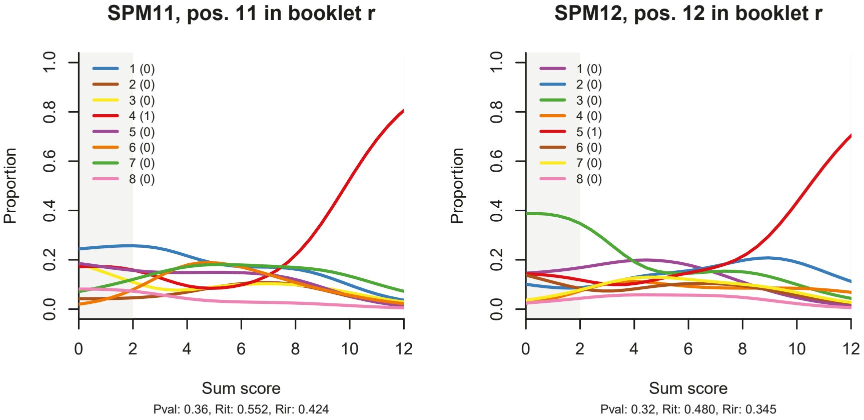


Figure 4. Non-parametric option-total regressions (distractor plots) for the twelve items in the SPM-LS test. The title of each plot shows the item label, in which booklet the item appears, and in what position. The legend shows the actual responses and the scores they will be given. Response alternatives that do not show up have not been chosen by any person.

Looking at the distractor plot for item SPM01, we observe that most of the seven distractors are not chosen at all, while one, represented with the blue line, is quite popular. When that happens, the item is effectively transformed into a coin tossing game. If this were a cognitive test, I would recommend rewriting the item. However, this is a matrix item, abstract and constructed in a principled manner, so the only explanation that comes to mind is that the test was given without a couple of items on which the examinees could train. For lack of these, the first item served as an exercise item.

Similar, but milder effects are observed on the ITR for items SPM09, SPM12, and possibly SPM11. The 3PL model (Appendix B) has larger pseudo-guessing parameters for these items. The distractor plots (Figure 4) show that all distractors are in use, but some of them tend to peak out in the middle, a bit like the middle category in a partial credit item. There might be some reason for this in the way Raven items are constructed.

There is more logic to guess when the item is difficult, especially if the stakes are high, but is this what is happening? Possibly. As one sharp-witted psychometrician told me once, the trouble with the research on guessing is that so much of it is guesswork. On the other hand, there must be ways to make guessing behaviour more explicit and include it in the rules of the game that we are playing. For example, one could have the subjects order the responses by their degree of confidence they are correct, or use a continuous response model as described in Verhelst (2019).

4. Discussion

To a pool of different analyses of the same dataset, I have contributed a specimen of the exploratory analysis we typically do when developing a cognitive test for assessment. My purpose was mainly to increase the diversity in professional perspectives, and to popularize some novel and useful diagnostic tools in our software.

While I use the Rasch model and the less popular interaction model, the focus is not on modelling, not even on the more traditional psychometric analysis of an intelligence test. Capitalizing on the fact that the models share the same scoring rule as the original test, the sum score, I use them to evaluate and support the scoring rule, and to highlight items that possibly go astray. I might have relied more heavily on the models in different circumstances: For example, if the test had more than one form (the Rasch model is useful for equating), or if I were interested in research, not in an instrument to assess persons.

The way in which I use the models explains why a paper that deals essentially with model fit does not treat model fit in the way typical of scientific research. I did not put forward any model to explain the data, in which case model fit would be an argument supporting my ideas. I did formulate a hypothesis or, rather, a guess (confirmed later) when I found out that a certain item did not follow my preferred model. In this case, model fit was about quality control more than about anything else.

I am certainly not original in pointing out that summative assessment, formative assessment, population surveys, methodological research and substantive research are sufficiently different to have not only distinct but sometimes even mutually exclusive criteria on what is desirable, appropriate, or admissible. This is fine as long as it is not forgotten and ignored.

In the final run, my story has three morals: (i) The way you should go “depends a good deal on where you want to get to” Carroll (1865), (ii) whatever the destination, always do exploratory analysis first, and (iii) in practical assessment, the model should follow from the scoring rule, not vice versa.

Funding: This research received no external funding.

Acknowledgments: I am greatly indebted to the other persons behind dexter: Gunter Maris and Timo Bechger at ACTNext, and Jesse Koops at Cito. Special thanks to the reviewers and to Timo Bechger and Norman D. Verhelst for many useful suggestions. Cito has graciously consented to cover the publication fee.

Conflicts of Interest: The author declares no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

2PL	Two-parameter logistic (model)
3PL	Three-parameter logistic (model)
CTT	Classical test theory
IM	Interaction model
IRF	Item response function
IRT	Item response theory
ITR	Item-total regression
PCM	Partial credit model
SPM-LS	Standard Progressive Matrices (last series)

Appendix A

This is a minimal script to perform all analyses in this paper. It does not cover the final formatting of the tables and graphs. Note that the original dataset was modified slightly by hand: column names were changed from STM1, STM2,. . . to STM01, STM02 etc.

```
library(dexter)                           # load the dexter library
setwd(’~/WD/Raven’)                       # set the work directory
keys = data.frame(                        # data frame as required
item_id = sprintf(’SPM%02d’, 1:12),     #   by keys_to_rules function
noptions = 8,
key = c(7,6,8,2,1,5,1,6,3,2,4,5)        #   (the correct responses)
)
rules = keys_to_rules(keys)               # scoring rules as reqd by dexter
```

```
db = start_new_project(rules, 'raven.db') # data base from the rules
dat = read.csv('dataset.csv', head=TRUE)  # read in data...
add_booklet(db, dat, 'r')                 # ... and add to the data base
tia_tables(db)                            # tables of CTT statistics
mo = fit_inter(db)                        # fit the Rasch and the IM
plot(mo)                                  # produce all ITR plots
distractor_plot(db,'SPM01')               # distractor plot for item SPM01
```

Appendix B

I tried to estimate the 3PL model for the SPM-LS dataset with three different programs: The R package mirt Chalmers (2012), the R package ltm Rizopoulos (2006), and the long-time flagship in educational testing, BILOG-MG Zimowski et al. (1996). All options in the R packages were held at their defaults, and no priors were used in BILOG-MG to constrain any of the three parameters during estimation. The results are shown in Table A1.

We observe reasonably good agreement between mirt and BILOG-MG, while the ltm estimates deviate more. Interesting enough, the estimates of the pseudo-guessing parameter, c, seem to agree the most among the three programs. They are also logical: since the items are arranged by increasing difficulty, there is no logic to guess on the easy items, so the guessing parameter is zero or close to zero. For the two most difficult items, it is close to 1/8, but it is difficult to explain why people should want to guess more on item 9. Moreover, all three programs seem to struggle with the discrimination parameter of item 11, quite seriously in the case of ltm.

Table A1. Parameter estimates for the 3PL model obtained for the SPM-LS dataset with three different programs.

Item	Mirt Estimates			Ltm Estimates			BILOG-MG Estimates		
	a	*b*	*c*	*a*	*b*	*c*	*a*	*b*	*c*
SPM01	0.85	−1.55	0.00	0.87	−1.51	0.00	0.83	−1.57	0.00
SPM02	1.93	−1.82	0.00	2.00	−1.76	0.00	2.00	−1.80	0.00
SPM03	1.61	−1.24	0.00	1.66	−1.21	0.00	1.62	−1.24	0.00
SPM04	3.65	−1.01	0.00	4.31	−0.95	0.00	3.60	−1.02	0.00
SPM05	4.70	−1.11	0.00	5.59	−1.04	0.00	4.57	−1.13	0.00
SPM06	2.26	−0.89	0.00	2.36	−0.86	0.00	2.23	−0.91	0.00
SPM07	1.55	−0.75	0.02	1.57	−0.76	0.00	1.55	−0.75	0.02
SPM08	1.58	−0.29	0.00	1.62	−0.29	0.00	1.57	−0.28	0.00
SPM09	2.28	0.19	0.24	2.27	0.18	0.23	2.27	0.19	0.24
SPM10	2.09	0.35	0.00	2.15	0.34	0.00	1.88	0.39	0.00
SPM11	5.83	0.63	0.11	32.28	0.67	0.12	6.04	0.63	0.11
SPM12	3.39	0.90	0.14	3.25	0.88	0.14	3.35	0.91	0.14

I made another comparison using only BILOG-MG and playing with the available priors for constraining parameter estimates. Table A2 shows results without any priors at all (same as in Table A1); with a lognormal (0, 0.5) prior on the discrimination parameter, a, and no prior on c; and with a lognormal (0, 0.5) prior on a and a beta$(20 \times \frac{1}{8} + 1, 20 \times \frac{7}{8} + 1)$ on c ($\frac{1}{8}$ and $\frac{7}{8}$ obtain from the fact that all items have 8 possible responses). The prior on a slightly alleviates the problem with the extremely high estimate while the prior on c simply invents guessing where we could not possibly have information on it: The less people have reason to guess, the more the estimate drifts towards $\frac{1}{8}$.

Table A2. Parameter estimates for the 3PL model obtained for the SPM-LS dataset with BILOG-MG and three different settings.

Item	Priors on a and c			Prior on a			No Prior		
	a	b	c	a	b	c	a	b	c
SPM01	0.90	−1.29	0.11	0.85	−1.53	0.00	0.83	−1.57	0.00
SPM02	1.93	−1.75	0.11	1.97	−1.80	0.00	2.00	−1.80	0.00
SPM03	1.65	−1.13	0.10	1.61	−1.24	0.00	1.62	−1.24	0.00
SPM04	3.23	−1.01	0.06	3.36	−1.03	0.00	3.60	−1.02	0.00
SPM05	3.85	−1.13	0.06	3.97	−1.15	0.00	4.57	−1.13	0.00
SPM06	2.34	−0.82	0.07	2.21	−0.90	0.00	2.23	−0.91	0.00
SPM07	1.64	−0.62	0.10	1.49	−0.80	0.00	1.55	−0.75	0.02
SPM08	1.67	−0.18	0.07	1.58	−0.29	0.00	1.57	−0.28	0.00
SPM09	1.79	0.05	0.16	1.91	0.10	0.19	2.27	0.19	0.24
SPM10	2.18	0.41	0.03	1.85	0.38	0.00	1.88	0.39	0.00
SPM11	3.97	0.64	0.10	3.98	0.63	0.10	6.04	0.63	0.11
SPM12	2.63	0.91	0.13	2.61	0.91	0.13	3.35	0.91	0.14

Appendix C

One of the reviewers has asked me to add some references on model fit at test level. Taken sufficiently seriously, this is not quite as easy at it may seem. Flowing out of the theory of errors, CTT is very concerned with test reliability and validity. Classical texts on CTT Gulliksen (1950) have entire chapters devoted to, say, the effect of test length on test error, reliability, and validity. IRT has an indisputable contribution in focusing on the item and item fit, but it may have gone a bit too far, overlooking the proverbial forest for the sake of the trees.

For the more pragmatic outlook of this paper, an important reference concerned with the practical implications of model misfit is Sinharay and Haberman (2014). Reasoning similar to mine, but pertaining to differential item functioning (DIF) rather than item fit, is found in Chalmers et al. (2016).

In what follows, I will try to avoid the item level—test level dichotomy, and steal a peek in-between. Our software, dexter Maris et al. (2019), has a handy function, `fit_domains()`, for the analysis of subtests within the test. The function transforms the items belonging to each subtest, or domain, into one large partial credit item. Such 'polytomisation', as discussed by Verhelst and Verstralen (2008), is a simple and efficient way to deal with testlets. The formal, constructed, and homogeneous nature of the SPM-LS test makes it a good candidate for some further experimentation. Note that I am not proposing a new method—I am just being curious.

I start by combining item 1, intended to be the easiest, with item 7, of medium difficulty. Item 2 will be combined with item 8, item 3 with item 9, and so on. We end up with 6 partial credit items combining an easier item with a more difficult one or, if you wish, six testlets or subtests made to be as parallel as possible. Their category trace lines are shown on Figure A1.

We can also examine the ITR (Figure A2), which are comparable to the ITR for the original test items; the item score on the y axis is also the 'test score' of the six subtests. We observe better item fit and a closer correspondence between the regressions predicted by the two models.

The next step will be to combine triplets of items: 1, 5, and 9; 2, 6, and 10, etc. Perhaps not surprisingly, the two models and the data come even closer (Figure A3).

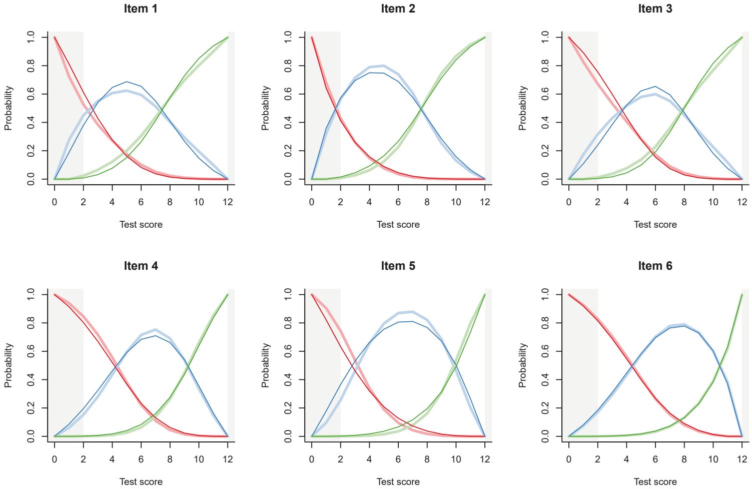


Figure A1. Category trace lines for partial credit items obtained by combining the original items SPM01 and SPM07 (Item 1), SPM02 and SPM08 (Item 2) etc. The partial credit model is shown with thinner and darker lines, and the polytomous IM with broader and lighter lines of the same hue.

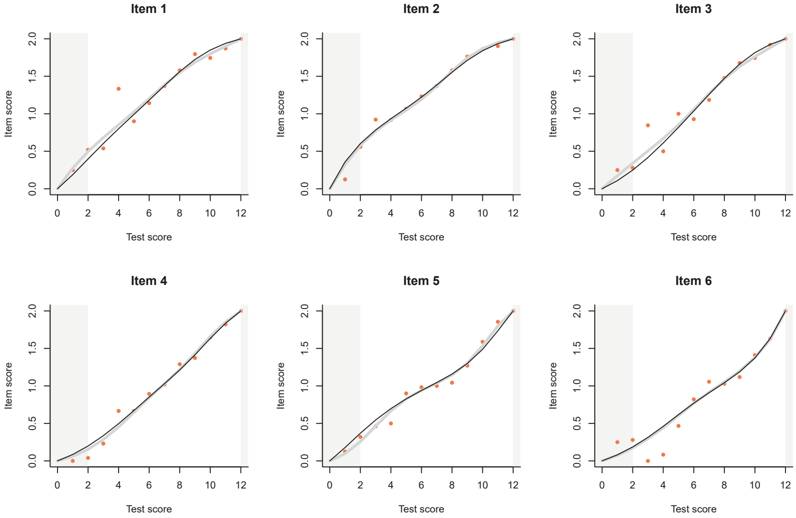


Figure A2. Item-total regressions for partial credit items obtained by combining the original items SPM01 and SPM07 (Item 1), SPM02 and SPM08 (Item 2) etc. Observed data is shown with pink dots, the PCM with thin black lines, and the interaction model with thick gray lines.

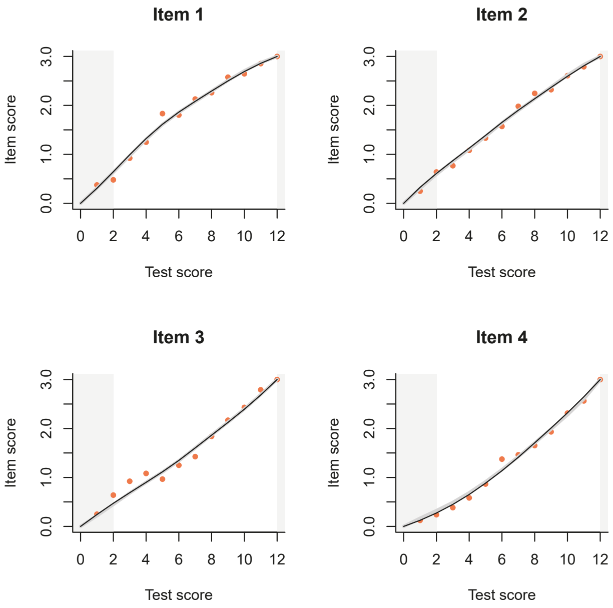


Figure A3. Item-total regressions for partial credit items obtained by combining triplets of items. Observed data is shown with pink dots, the PCM with thin black lines, and the interaction model with thick gray lines.

Quite predictably, the next step is to combine quadruples of items, and the result is even better fit (Figure A4).

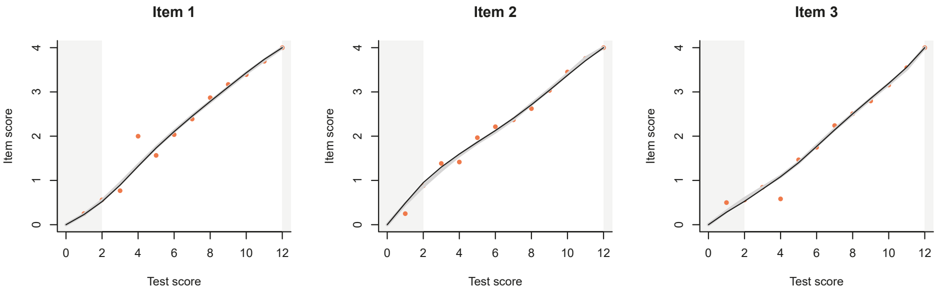


Figure A4. Item-total regressions for partial credit items obtained by combining quadruples of items. Observed data is shown with pink dots, the PCM with thin black lines, and the interaction model with thick gray lines.

Finally, we have two subtests of six items each, one consisting of the odd-numbered items, and the other of the even-numbered items in the original test (Figure A5). This parcours is the closest approximation to item fit at test level from an IRT perspective that I can produce as of now.

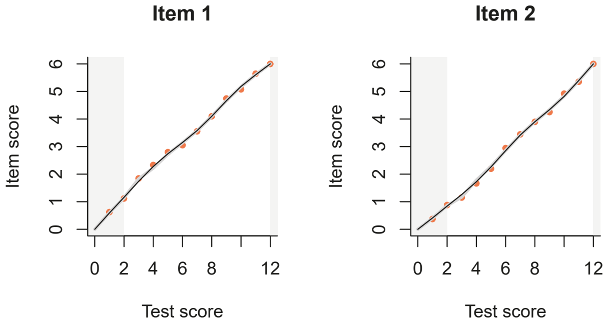


Figure A5. Item-total regressions for two subtests of six items each. Observed data is shown with pink dots, the PCM with thin black lines, and the interaction model with thick gray lines.

References

American Psychological Association. 2010. *Publication manual of the American Psychological Association*, 6th ed. Washington: American Psychological Association.

Andersen, Erling B. 1973.Conditional inference for multiple-choice questionnaires. *British Journal of Mathematical and Statistical Psychology* 26: 31–44. [CrossRef]

ATP Tour Inc. 2020. *The 2020 ATP® Official Rulebook*. Available online: https://www.atptour.com/en/corporate/rulebook (accessed on 1 April 2020).

Azevedo, C. L. N. 2009. Some Observations on the Identification and Interpretation of the 3PL IRT Model. *Measurement: Interdisciplinary Research and Perspectives* 7: 89–91. [CrossRef]

Bock, R. Darrell, and Murray Aitkin. 1981. Marginal Maximum Likelihood Estimation of Item Parameters: Application of an EM Algorithm. *Psychometrika* 46: 443–59. [CrossRef]

Bock, R. Darrell. 1972. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika* 37: 29–51. [CrossRef]

Brouwers, S. A., F. J. van de Vijver, and D. A. van Hemert. 2009. Variation in Raven's Progressive Matrices scores across time and place. *Learning and Individual Differences* 19: 330–38. [CrossRef]

Carroll, Lewis. 1865. *Alice's Adventures in Wonderland*. London: MacMillan.

Chalmers, Robert P. 2012. mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software* 48: 1–29. [CrossRef]

Chalmers, Robert P., A. Counsell, and D. B. Flora. 2016. It Might Not Make a Big DIF: Improved Differential Test Functioning Statistics That Account for Sampling Variability. *Educational and Psychological Measurement* 76: 114–40. [CrossRef]

Dorans, Neil J. 2012. The Contestant Perspective on Taking Tests: Emanations From the Statue Within. *Educational Measurement: Issues and Practice* 31: 20–37. [CrossRef]

Garcia-Garzon, Eduardo, Francisco J. Abad, and Luis E. Garrido 2019. Searching for G: A New Evaluation of SPM-LS Dimensionality. *Journal of Intelligence* 7: 14.

Glas, Cees A. W. 2009. What IRT Can and Cannot Do. *Measurement: Interdisciplinary Research and Perspectives* 7: 91–93. [CrossRef]

González, Jorge, and Marie Wiberg. 2017. *Applying Test Equating Methods: Using R*. Berlin/Heidelberg: Springer. [CrossRef]

Gulliksen, Harold. 1950. *Theory of Mental Tests*. Hoboken: Wiley.

Haberman, Shelby J. 2007. The Interaction Model. In *Multivariate and Mixture Distribution Rasch Models: Extensions and Applications*. Edited by M. von Davier and C. H. Carstensen. New York: Springer, chap. 13, pp. 201–16.

Kolen, Michael J., and Robert L. Brennan. 2014. *Test Equating, Scaling, and Linking: Methods and Practices*, 3rd ed.; New York: Springer. [CrossRef]

Koops, Jesse, Eva de Schipper, Ivailo Partchev, Gunter Maris, and Timo Bechger. 2019. *dextergui: A Graphical User Interface for Dexter* (Version 0.2.0). R Package. Available online: https://cran-r.project.org (accessed on 1 April 2020).

Leighton, J. P., and M. J. E. Gierl. 2007. *Cognitive Diagnostic Assessment for Education: Theory and Applications*. Cambridge: Cambridge University Press. [CrossRef]

Lord, Frederic M. 1980. *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale: Lawrence Erlbaum.

Lord, Frederic M., and Melvin R. Novick. 1968. *Statistical Theories of Mental Test Scores (with Contributions by A. Birnbaum)*. Reading: Addison-Wesley.

Maris, Gunter, and Timo Bechger. 2009. On Interpreting the Model Parameters for the Three Parameter Logistic Model. *Measurement: Interdisciplinary Research and Perspectives* 7: 75–88. [CrossRef]

Maris, Gunter, Timo Bechger, Jesse Koops, and Ivailo Partchev. 2019. *Dexter: Data Management and Analysis of Tests* (Version 1.0.1). R Package. Available online: https://cran-r.project.org (accessed on 1 April 2020).

Masters, Geoffrey N. 1982. A Rasch Model for Partial Credit Scoring. *Psychometrika* 47: 149–74. [CrossRef]

Myszkowski, Neil, and M. Storme. 2018. A snapshot of g? Binary and polytomous item-response theory investigations of the last series of the Standard Progressive Matrices (SPM-LS). *Intelligence* 68: 109–16. [CrossRef]

Partchev, Ivailo. 2009. 3PL: A Useful Model with a Mild Estimation Problem. *Measurement: Interdisciplinary Research and Perspectives* 7: 94–96. [CrossRef]

R Core Team. 2013. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Available online: http://www.R-project.org/ (accessed on 1 April 2020).

Rasch, Georg. 1980. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago: University of Chicago Press. First published 1960.

Raven, J. C. 1941. Standardization of Progressive Matrices, 1938. *British Journal of Medical Psychology* 19: 137–50. [CrossRef]

Rizopoulos, Dimitris. 2006. ltm: An R package for Latent Variable Modelling and Item Response Theory Analyses. *Journal of Statistical Software* 17: 1–25. [CrossRef]

San Martín, Ernesto, Jorge González, and Francis Tuerlinckx. 2009. Identified Parameters, Parameters of Interest and Their Relationships. *Measurement: Interdisciplinary Research and Perspectives* 7: 97–105. [CrossRef]

Sinharay, Sandip, and Shelby J. Haberman. 2014. How Often Is the Misfit of Item Response Theory Models Practically Significant? *Educational Measurement: Issues and Practice* 33: 23–35. [CrossRef]

Thissen, David. 1976. Information in Wrong Responses to the Raven Progressive Matrices. *Journal of Educational Measurement* 13: 201–14. [CrossRef]

Thissen, David. 2009. On Interpreting the Parameters for any Item Response Model. *Measurement: Interdisciplinary Research and Perspectives* 7: 106–10. [CrossRef]

Verhelst, Norman D. 2019. Exponential Family Models for Continuous Responses. In *Theoretical and Practical Advances in Computer-Based Educational Measurement*. Edited by B. P. Veldkamp and C. Sluijter. Berlin/Heidelberg: Springer, chap. 7, pp. 135–59.

Verhelst, Norman D., and Huub Verstralen. 2008. Some Considerations on the Partial Credit Model. *Psicologica: International Journal of Methodology and Experimental Psychology* 29: 229–54.

von Davier, Matthias. 2009. Is There Need for the 3PL Model? Guess What? *Measurement: Interdisciplinary Research and Perspectives* 7: 110–14. [CrossRef] von Davier, Alina, ed. 2011. *Statistical Models for Test Equating, Scaling, and Linking*. Berlin/Heidelberg: Springer. [CrossRef]

von Davier, Alina, Paul W. Holland, and Dorothy T. Thayer. 2004. *The Kernel Method of Test Equating*; Berlin/Heidelberg: Springer. [CrossRef]

Zimowski, Michelle. F., Eiji Muraki, Rober J. Mislevy, and R. Darrell Bock. 1996. *BILOG–MG. Multiple-Group IRT Analysis and Test Maintenance for Binary Items*. Chicago: SSI Scientific Software International.

Journal of
Intelligence

Article

A Mokken Scale Analysis of the Last Series of the Standard Progressive Matrices (SPM-LS)

Nils Myszkowski

Department of Psychology, Pace University, New York, NY 10038, USA; nmyszkowski@pace.edu

Received: 10 January 2020; Accepted: 24 April 2020; Published: 6 May 2020

Abstract: Raven's Standard Progressive Matrices (Raven 1941) is a widely used 60-item long measure of general mental ability. It was recently suggested that, for situations where taking this test is too time consuming, a shorter version, comprised of only the last series of the Standard Progressive Matrices (Myszkowski and Storme 2018) could be used, while preserving satisfactory psychometric properties (Garcia-Garzon et al. 2019; Myszkowski and Storme 2018). In this study, I argue, however, that some psychometric properties have been left aside by previous investigations. As part of this special issue on the reinvestigation of Myszkowski and Storme's dataset, I propose to use the non-parametric Item Response Theory framework of Mokken Scale Analysis (Mokken 1971, 1997) and its current developments (Sijtsma and van der Ark 2017) to shed new light on the SPM-LS. Extending previous findings, this investigation indicated that the SPM-LS had satisfactory scalability ($H = 0.469$), local independence and reliability ($MS = 0.841$, $LCRC = 0.874$). Further, all item response functions were monotonically increasing, and there was overall evidence for invariant item ordering ($H_T = 0.475$), supporting the Double Monotonicity Model (Mokken 1997). Item 1, however, appeared problematic in most analyses. I discuss the implications of these results, notably regarding whether to discard item 1, whether the SPM-LS sum scores can confidently be used to order persons, and whether the invariant item ordering of the SPM-LS allows to use a stopping rule to further shorten test administration.

Keywords: Mokken scale analysis; non-parametric item response theory; psychometrics; invariant item ordering

1. Introduction

The general factor of intelligence (g) is central in the prediction of several outcomes, such as job performance (Ree and Earles 1992; Salgado et al. 2003) and academic achievement (Rohde and Thompson 2007). Its accurate measurement is therefore crucial in multiple contexts, including personnel selection, vocational guidance or academic research in individual differences. However, because of practical constraints, it is desirable in many contexts to reduce test length as much as possible, while maintaining acceptable accuracy.

Raven's Standard Progressive Matrices (SPM) (Raven 1941) and Advanced Progressive Matrices (APM) (Raven et al. 1962) are widely used—though also criticized (Gignac 2015)—instruments to measure g. However, both these tests remain rather long in untimed conditions, with some participants sometimes taking more than 40 min to respond them (Hamel and Schmittmann 2006). Several solutions have been proposed to further reduce test administration time, such as constraining time (Hamel and Schmittmann 2006) and using short versions (Bors and Stokes 1998).

While these solutions have focused on the APM (Hamel and Schmittmann 2006; Bors and Stokes 1998; Myszkowski and Storme 2018) have recently suggested that the last series of the SPM—the SPM-LS—could be a more efficient solution, with only 12 items, while maintaining the progressive aspect characteristic of Raven's matrices, along with satisfactory psychometric properties. However, the original study (Myszkowski and Storme 2018)—which I propose to extend—has studied the SPM-LS with parametric Item Response Theory (IRT) models, and is largely focused on recovering information from distractor responses using nested logit models (Suh and Bolt 2010; Storme et al. 2019), therefore putting aside important aspects of the test—such as the monotonicity of item responses and invariant item ordering, which I later further discuss. I propose here to bridge these gaps using the framework of Mokken Scale Analysis (MSA) (Mokken 1971, 1982, 1997), a well developed non-parametric item-response theory framework that is particularly appropriate to address them (Sijtsma and van der Ark 2017).

1.1. The SPM-LS

While the SPM is heavily studied, the SPM-LS is very recent, and thus has not been the object of many investigations. Currently, it has only been studied in its original investigation (Myszkowski and Storme 2018)—which used binary and nominal IRT models—and as part of this special issue through a further investigation of its dimensionality (Garcia-Garzon et al. 2019). Investigations of the SPM-LS indicated that IRT models could satisfactorily fit test responses (Bürkner 2020; Myszkowski and Storme 2018), and that the test seemed to present adequate reliability/information for abilities ranging from about 2 standard deviations below the mean—or 3 if recovering information from distractors—to 1.5 to 2 standard deviations above the mean (Myszkowski and Storme 2018), in a sample of undergraduate students, suggesting that it could be more appropriate in terms of difficulty for the general population than for post-secondary students. In addition, Garcia-Garzon et al. (2019) notably studied in this special issue the dimensionality of the SPM-LS using a variety of methods—Exploratory Graph Analysis (EGA), bifactor Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA). Overall, the psychometric qualities of the SPM-LS so far appeared satisfactory for use by researchers and practitioners, but some characteristics have not been studied, for which Mokken Scale Analysis is a particularly appropriate framework.

1.2. Mokken Scale Analysis

Since its inception (Mokken 1971), Mokken Scale Analysis has been the object of several methodological developments, notably discussing how to evaluate the properties of instruments evaluated with MSA (Van der Ark 2012), best practices in MSA (Sijtsma and van der Ark 2017) and the active development of a package (Van der Ark 2007) for the statistical programming language R. While it is largely and more thoroughly described elsewhere (Mokken and Lewis 1982; Mokken 1997; Van der Ark 2007; Sijtsma and van der Ark 2017; Sijtsma 1998), I could briefly describe Mokken Scale Analysis (MSA) (Mokken and Lewis 1982; Mokken 1997) as a non-parametric IRT framework, which, for dichotomous responses, represents the probability of succeeding an item j as a function of an person i's latent ability—θ_i. Unlike the Rasch model (Rasch 1993) and, more broadly, unlike binary logistic and normal ogive models—which are said to be parametric IRT models (Mokken and Lewis 1982)—MSA does not represent the relation between latent ability and item responses using item parameters, but using an item-response function only defined as monotonically increasing (Mokken and Lewis 1982).

1.3. The Benefits of Mokken Scale Analysis

Because they do not require response functions to have a specific shape, Mokken's models are less constrained than (notably) Rasch models (Meijer et al. 1990), which implies that some items that are not well fitted by Rasch models may still be scalable with MSA, because their response function may be monotonic without necessarily having a logistic/normal ogive shape. While MSA does not allow certain applications otherwise permitted by Rasch modeling, like test equating or computer adaptive testing, (Meijer et al. 1990, p. 297) note that, "for many testing applications, it often suffices to know the order of persons on an attribute". Therefore, Mokken scaling is attractive for the reason that it focuses mainly on a test's capacity to order persons, while allowing for more items to fulfill its requirements than Rasch models do allow. In the context of the SPM-LS, this is particularly interesting, especially as Myszkowski and Storme (2018) had to use highly parametrized models to achieve an acceptable fit, with 3- and 4-parameter models fitting much better than notably the Rasch 1-parameter model—in this special issue, Bürkner (2020) makes a similar conclusion using Bayesian IRT. Instead of increasing the number of parameters to better fit item responses—and risking overfitting and thus compromising reproducibility—Mokken scaling proposes to retain fewer (but fundamental) uses of a test: Ordering persons (for both MSA models) and items (for the Double Monotonicity model only).

1.3.1. The Monotone Homogeneity and Double Monotonicity Models

For dichotomous items, Mokken (1997) defined two item-response models: The Monotone Homogeneity Model (MHM) and the Double Monotonicity Model (DMM). Both the Monotone Homogeneity Model and the Double Monotonicity Model assume the monotonicity of item response functions. However, the two models differ in that only the Double Monotonicity Model assumes that item response functions do not intersect—an assumption usually referred to as invariant item ordering.

Before focusing on these two assumptions central to MSA, as well as their consequences in the context of the SPM-LS, it is important to note that both models also assume unidimensionality, meaning that they both assume that the same latent attribute θ_i explains the items scores—therefore also assuming local independence (Sijtsma et al. 2011). While MSA offers procedures (also used in this study) to investigate this assumption, they would probably not justify a new study, because the dimensionality of the SPM-LS has been, on this very dataset, investigated with a plethora of psychometric methods (Myszkowski and Storme 2018; Garcia-Garzon et al. 2019). I will therefore mainly focus here on the *incremental* value of using Mokken Scale Analysis in addition to these previously used approaches.

1.3.2. Monotonicity of Item Response Functions

An important feature of MSA is that it allows to study monotonicity, where parametric Item-Response Theory and traditional (linear) factor analysis models generally leave this assumption untested. Indeed, although parametric item response models for binary responses are (in general) monotonous, a misfitting item does not necessarily indicate that the item response is non-monotonic (Meijer et al. 1990). Therefore, because it has only been studied with parametric response models, the monotonicity of the SPM-LS has remained untested so far. This characteristic is manifested, in pass-fail (binary) tests like the SPM-LS, by item response functions that are monotonically increasing. This means that the probability to succeed on an item monotonically increases with the examinee's ability. In contrast with parametric IRT, the framework of Mokken Scale Analysis offers methods to investigate this property and specifically identify its violations (Van der Ark 2007). I therefore propose, in the present study, to use this framework to bridge that gap in the study of the test.

As a consequence, this study is therefore the first to study the monotonicity of the SPM-LS, which, as previously noted (Van der Ark 2007), is not only relevant to Mokken scaling, but also relevant to any model that formulates this assumption, such as parametric Item-Response Theory models and traditional factor analysis. It is an essential psychometric property of a test, because it is the property that implies that higher scores imply higher abilities (for all items, at any ability level), and thus that scores can be used to infer person ordering (Van der Ark 2007).

1.3.3. Invariant Item Ordering

While both the Monotone Homogeneity Model and the Double Monotonicity Model assume unidimensionality and monotonicity of the response functions, only the Double Monotonicity Model assumes that the ordering of the items (based on their difficulty) is the same for all examinees (Mokken 1997; Sijtsma and van der Ark 2017; Sijtsma et al. 2011). In other words, this property, referred to as Invariant Item Ordering (IIO), assumes that, for any given item pair, the easier item has a higher probability of being succeeded than the more difficult one at any ability level. This manifests itself graphically by the item response functions of the two items not intersecting.

As was previously noted (Sijtsma et al. 2011; Sijtsma and van der Ark 2017), this property is an important feature of a test, as it "greatly facilitates the interpretation of test scores" (Sijtsma and van der Ark 2017, p. 143), and is "both omnipresent and implicit in the application of many tests, questionnaires, and inventories" (Ligtvoet et al. 2010, p. 593). Indeed, a stronger IIO implies that two persons with the same total score are more likely to have succeeded the same items, and that an examinee with a higher total score than another examinee is more likely to have answered correctly the same items, and one or several more difficult items. Therefore, invariant item ordering lends more meaning to person comparisons based on total scores.

In addition, IIO is especially relevant for the SPM-LS, because its items substantially vary in difficulty and are presented by increasing difficulty. A stronger IIO implies that, if an examinee fails an item, there is an increased probability that the examinee will fail the next (more difficult) one. Therefore, a stronger IIO would suggest that we can envision stopping the test administration after one or several items have been failed (Sijtsma and Meijer 1992). This would presents practical advantages, notably for shortening test administration.

2. Materials and Methods

2.1. Participants

Per the topic of this special issue, I re-analyzed the publicly available dataset from Myszkowski and Storme (2018) study. The original study presented various parametric IRT analyses performed on a dataset comprised of 499 students (214 males and 285 females) aged between 19 and 24. Because I directly reanalysed this dataset, I point to the original article for more details on data collection and sample characteristics.

One thing to note that is specific to this paper is that the sample size is both similar to the one used in Sijtsma and van des Ark's tutorial on Mokken scale analysis (Sijtsma and van der Ark 2017) and, more importantly, in accordance with the sample size recommendations provided by Straat et al. (2014). They show (p. 817) that a sample size of around 500 is largely sufficient for an accurate analysis with scalability coefficients H_j of 0.42 (or higher)—in the results, I present scalability coefficients, and show that the scalability of the scale meets that requirement.

2.2. Instrument

The Last Series of the Standard Progressive Matrices, or SPM-LS (Myszkowski and Storme 2018), was built from the original Standard Progressive Matrices (Raven 1941), a very popular and extensively researched test of non-verbal logical reasoning, which is also frequently used as a brief measure of *g*, the general factor of intelligence. As its name indicates, it consists of the last—and thus most difficult—series of the original SPM, but used as a standalone test (without examinees taking previously the other series). It is composed of 12 items of theoretically increasing difficulty. Each item consists of an incomplete 3-by-3 matrix, with the last element of the matrix being missing. The examinee is to identify, among eight options—seven distractors and one correct response—the missing matrix element.

Research shows that *g* is far from being extensively, nor purely captured by the SPM (Gignac 2015; Carpenter et al. 1990), and this is certainly even more true of SPM-LS, since it is a shortened version. Nevertheless, the SPM, and a fortiori the SPM-LS, present the advantage of being short measures, with overall satisfactory reliability. In particular, the SPM-LS, in its original investigation on this dataset (Myszkowski and Storme 2018), presented encouraging evidence of reliability, with observed reliabilities based on IRT modeling that ranged from 0.78 to 0.84 depending on the IRT model used, and a Cronbach's α of 0.92.

As unidimensionality is an assumption of Mokken Scale Analysis (Van der Ark 2007), it is also important to note that the SPM-LS investigations indicated that the test is essentially unidimensional, with a McDonald's coefficient ω_h of 0.86 (Myszkowski and Storme 2018) and satisfactory fit of unidimensional models (Myszkowski and Storme 2018). Garcia-Garzon et al. (2019) explorations also supported unidimensionality, in spite of a nuisance factor specific to the last six items.

2.3. Analysis

Because Sijtsma and van der Ark's tutorial on Mokken scale analysis (Sijtsma and van der Ark 2017) presents the advantages of presenting the current state of the art of Mokken scale analysis and of laying out clearly the different steps to take in order to perform a Mokken scale analysis, I followed the different steps provided in the tutorial. All analyses were computed using the same team's regularly updated and comprehensive R package `mokken` (Van der Ark 2007, 2012; Sijtsma et al. 2011) (version 2.8.11).

A reason for the popularity of Mokken scaling is the availability of an automatic procedure to select a set (or several sets) of scalable items, a procedure generally referred to as the Automated Item Selection Procedure (AISP), which aims at maximizing scalability. In addition, Straat et al. (2016) also recently suggested a item selection procedure which aims to maximize local independence. Likewise, a stepwise selection procedure aiming at maximizing invariant item ordering has been proposed (Ligtvoet et al. 2010). Still, it was decided here that the primary objective of the present study would be to investigate the SPM-LS as an a priori scale, meaning that the main objective was to investigate its qualities using Mokken Scale Analysis, not to carve a revised instrument out of it. This decision was motivated by the fact that the SPM-LS is already a very short measure (12 items), and also because, in the SPM-LS, the very process of solving items is—at least theoretically—used to help the examinee learn the rule(s) used in subsequent items (Myszkowski and Storme 2018). Therefore, even if an item were to present poor qualities (e.g., weak scalability), it might still be useful as a training for the other items, and thus it may still be preferable or conservative to keep it.

2.3.1. Data Preparation

The dataset analyzed did not present any missing data nor impossible responses. Sijtsma and van der Ark (2017) recommend, as a preliminary step to Mokken Scale Analysis, to filter out cases whose responses are dubious, and they suggest doing so using the count of Guttman errors. I proceeded to count the number of Guttman errors G_+ per case, computed with the package function `check.errors()` of the `mokken` package. There were a total of 2021 Guttman errors, indicating that the items did not constitute a Guttman scale.

As Sijtsma and van der Ark (2017) suggested, I identified as dubious cases—and consequently removed—the cases for which G_+ indices were beyond the upper Tukey fence of the distribution of G_+ indices. This corresponded to cases with more than 15 Guttman errors, and resulted in the elimination of 14 cases (1.17% cases) with suspicious item-score patterns. The frequency histogram of G_+ indices, with the Tukey fence, is presented in Figure 1.

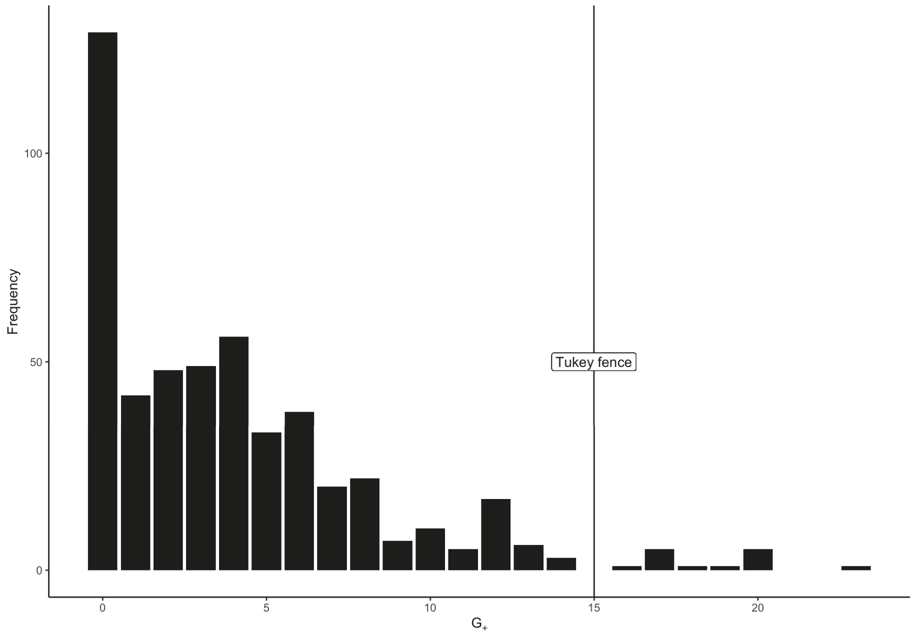


Figure 1. Histogram of the count of Guttman errors (G_+), with Tukey fence (3rd quartile $+1.5 \times IQR$) used as a threshold for outlier detection.

2.3.2. Scalability

As recommended by Sijtsma and van der Ark (2017), I investigated the scalability of the complete SPM-LS by computing H_{jk} (scalability coefficients for item pairs), H_j (scalability coefficients for items) and H (total scalability coefficient of the scale). I used the rules of thumb originally proposed by Mokken (1997) and currently suggested by Sijtsma and van der Ark (2017), which are $H < 0.3$ for insufficient scalability, $0.3 \leq H < 0.4$ for weak scalability, $0.4 \leq H < 0.5$ for medium scalability and $H \geq 5$ for strong scalability. Since the Monotone Homogeneity Model implies that H_{jk} and H_j are all positive (and ideally as close to 1 as possible), I searched for negative values (or values close to 0) as violations of the monotonicity (Sijtsma and van der Ark 2017).

2.3.3. Local Independence

Local independence is an assumption of both the monotone homogeneity model and the double monotonicity item. Local independence implies that item scores are independent for a given ability level θ. As suggested in Sijtsma and van der Ark (2017)'s tutorial, I used the procedure proposed by Straat et al. (2016) to study local dependencies in the SPM-LS. They suggest the computation of three series of indices: W_1, W_2 and W_3. While the computation of these indices is further explained in the original article, we can note that high W_1, W_2 and W_3 values indicate local dependencies. High W_1 values indicate that an item pair is likely positively locally dependent. An item with a high W_2 is likely to be positively locally dependent with another item. High W_3 indicate that an item pair is likely negatively locally dependent. Again here, and as Straat et al. (2016) suggested, a Tukey fence was used to detect problematic items.

2.3.4. Monotonicity

As recommended by Sijtsma and van der Ark (2017), I studied monotonicity by plotting item response functions, using a non-parametric regression of each item scores on "rest scores" (the total scores on the other items) (Junker and Sijtsma 2000). Following the defaults of the `check.monotonicity()` function of the `mokken` package (Van der Ark 2007), the rest scores were grouped using a minimum size of $N/5$ (meaning groups of 100 cases at least). The identified violations were then significance tested (Van der Ark 2007; Molenaar and Sijtsma 2000).

As an alternative to testing monotonicity violations, it has been been recently proposed that positive evidence for monotonicity can be gathered through Bayes factors (Tijmstra et al. 2015; Tijmstra and Bolsinova 2019). Based on a suggestion by a reviewer that I use this procedure, I contacted the first author of these papers, who provided code to implement it. The procedure is discussed in more details in the original paper (Tijmstra et al. 2015), but, in short, it consists in evaluating the relative amount of support from the data for (strict) manifest monotonicity—denoted hypothesis H_{MM}—against the competing hypothesis that there is at least one manifest non-monotonicity—denoted hypothesis H_{NM}—and against the competing hypothesis of essential monotonicity—denoted H_{EM}, defined as a form of monotonicity that allows for non-monotonicities between adjacent manifest scores. Bayes Factors $BF_{MM,NM}$ and $BF_{MM,EM}$ were estimated through Gibbs sampling, and used to indicate support for H_{MM} in contrast with H_{NM} and H_{EM} respectively. Values above 1 indicate more support for H_{MM} than for the competing hypothesis. 20,000 iterations were used as burn-in and discarded, and 100,000 iterations were subsequently used to estimate the Bayes Factors—which is more conservative than initially suggested (Tijmstra et al. 2015).

2.3.5. Invariant Item Ordering

Even though Invariant Item Ordering (IIO) is only an assumption of the Double Monotonicity Model for binary items (Mokken 1997)—not of the the Monotone Homogeneity Model—I studied IIO because of its benefits for score interpretability and the possibility to stop the examination after failed items. As Sijtsma and van der Ark (2017) suggested, overall IIO was assessed with the coefficient H_T. Like for the H scalability coefficients, and as suggested by Ligtvoet et al. (2010), I used thresholds of $H_T < 0.3$ for insufficient IIO (Sijtsma and Meijer 1992), $0.3 \leq H_T < 0.4$ for weak IIO, $0.4 \leq H_T < 0.5$ for medium IIO and $H_T \geq 5$ for strong IIO. I also graphically compared the item response functions of pairs of items that significantly intersect.

2.3.6. Reliability

Cronbach's α and empirical reliability from parametric IRT models have been previously reported and discussed as satisfactory in the same dataset (Myszkowski and Storme 2018). Here, to investigate reliability, as recommended in the context of MSA (Sijtsma and van der Ark 2017), I used the Molenaar-Sijtsma (MS) reliability estimate (Sijtsma and Molenaar 1987), which assumes the Double Monotonicity Model. In addition, I reported the Latent Class Reliability Coefficient (LCRC) (van der Ark et al. 2011), which is more robust to violations of the Double Monotonicity Model.

3. Results

3.1. Scalability

The SPM-LS had medium scalability, with an H coefficient of 0.469 ($SE = 0.021$). The scalability of the item pairs H_{jk} is reported in Table 1, along with the scalability of the items H_j. All item pairs and item scalability coefficients were positive, giving support to the monotone homogeneity model. However, it can be noted that the first item had a substantially lower scalability ($H_j = 0.265$) than the other items (H_j ranging from 0.401 to 0.602), and that the total scalability would be strong ($H = 0.516$) without this item.

Table 1. Scalability coefficients of the item pairs (H_jk) and items (H_j).

Index	Item	1	2	3	4	5	6	7	8	9	10	11	12
H_{jk}	1												
	2	0.616											
	3	0.331	0.535										
	4	0.248	0.613	0.286									
	5	0.294	0.493	0.421	0.772								
	6	0.272	0.511	0.509	0.520	0.675							
	7	0.122	0.544	0.295	0.471	0.575	0.362						
	8	0.263	0.647	0.442	0.636	0.701	0.547	0.429					
	9	0.095	0.393	0.460	0.441	0.481	0.416	0.427	0.378				
	10	0.327	0.709	0.799	0.938	0.921	0.743	0.526	0.449	0.403			
	11	0.399	0.664	0.569	0.822	0.774	0.677	0.595	0.506	0.522	0.467		
	12	0.267	0.717	0.253	0.839	0.847	0.576	0.613	0.602	0.462	0.400	0.449	
H_j		0.265	0.568	0.426	0.545	0.602	0.499	0.422	0.476	0.401	0.529	0.536	0.500

3.2. Local Independence

The W_1, W_2 and W_3 indices for local dependencies detection are presented in Table 2. While W_2 indices did not suggest that any item were likely in a positive locally dependent pair, W_1 indices identified 3 positive local dependencies, between Item 4 and 11, 5 and 11, and 5 and 12. W_3 indices suggested a negative local dependency between item 1 and 9.

3.3. Monotonicity

The item response functions of all items are presented in Figure 2. Only one violation of monotonicity was observed, for item 3—the response function of this item can be seen as slightly decreasing between rest scores 8–9 and 10–11. This violation was, however, non significant.

The Bayes Factors used to compare the relative support for monotonicity against non-monotonicity and essential monotonicity are reported in Table 3. Overall, monotonicity was supported for all items against its complement—although the support was much weaker for the last two items—with Bayes Factors ranging from 1.64 to 818,417.9. The data tended to support (strict) monotonicity against essential monotonicity, with, however, limited support, and with the exception of 12, which had a Bayes Factor slightly smaller than 1.

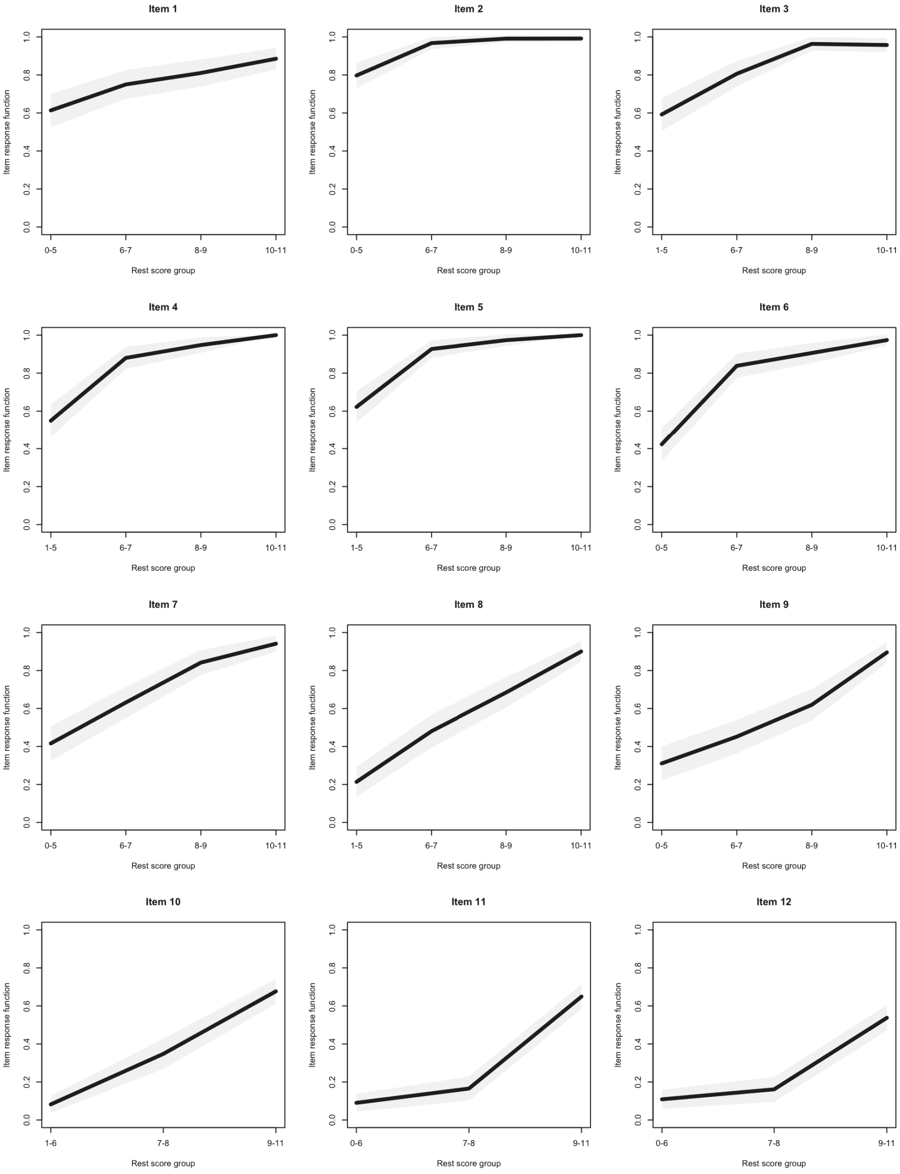


Figure 2. Item response functions of the last series of the Standard Progressive Matrices (SPM-LS) items (with 95% confidence intervals).

Table 2. W_1, W_2 and W_3 indices of the SPM-LS (flagged values in bold face).

Index	Item	1	2	3	4	5	6	7	8	9	10	11	12
W_1	1		3.044	3.621	3.842	3.852	4.714	2.897	3.963	1.857	2.956	1.955	1.556
	2	0.508		1.865	3.742	3.196	1.908	1.497	2.806	0.661	1.932	0.349	0.309
	3	1.073	2.381		2.512	3.065	2.503	1.324	2.432	1.449	3.655	0.909	0.634
	4	0.036	2.128	0.503		3.039	1.018	0.468	1.772	0.095	2.959	0.019	0.162
	5	0.072	1.844	0.430	3.299		1.043	0.591	2.739	0.181	2.915	0.021	0.037
	6	0.352	1.996	0.389	1.668	2.098		0.392	1.237	0.159	2.578	0.101	0.704
	7	0.766	2.287	1.529	2.730	3.621	1.666		1.523	0.779	2.265	0.282	0.310
	8	0.136	1.450	0.445	2.325	2.537	0.721	0.775		0.502	0.605	0.045	1.769
	9	0.483	3.996	2.731	3.375	3.486	2.376	2.507	3.225		2.176	1.241	0.873
	10	0.077	3.838	1.735	3.742	3.316	1.245	0.893	0.317	0.326		0.448	0.107
	11	0.425	5.611	1.779	**8.765**	**8.429**	2.972	1.813	0.854	1.451	0.994		0.274
	12	1.137	5.129	1.017	5.455	**7.380**	3.288	2.525	2.370	2.441	2.234	2.089	
W_2		49.281	38.952	42.890	30.910	27.740	39.323	44.227	33.246	44.265	28.471	35.393	33.611
W_3	1												
	2	3.116											
	3	3.487	2.708										
	4	4.338	3.408	5.276									
	5	3.826	3.321	3.457	0.297								
	6	4.534	5.199	1.561	2.944	2.022							
	7	6.683	3.861	5.579	2.820	1.616	5.398						
	8	3.869	3.653	4.941	2.959	2.919	3.122	2.187					
	9	**7.181**	5.033	3.288	4.118	4.116	5.143	4.575	1.376				
	10	4.405	3.469	2.120	0.990	1.626	2.111	3.628	3.555	2.584			
	11	3.540	2.756	4.269	2.037	2.424	3.074	5.086	3.468	3.870	1.604		
	12	4.303	2.428	6.204	1.723	2.116	4.215	2.797	1.198	2.982	2.378	3.267	

Table 3. Bayes Factor for the relative support of monotonicity against its complement ($BF_{MM,NM}$) and against essential monotonicity ($BF_{MM,EM}$).

Item	$BF_{MM,NM}$	$BF_{MM,EM}$
1	825.15	1.44
2	34,666.90	3.49
3	57,682.66	3.89
4	871,824.00	8.26
5	89,668.37	4.20
6	95.22	1.15
7	9594.47	4.13
8	818,417.90	6.40
9	12.08	2.58
10	50,455.13	3.80
11	1.98	4.31
12	1.64	0.81

3.4. Invariant Item Ordering

The observed invariant item ordering was medium but close to strong, with a H_T coefficient of 0.475, overall supporting IIO, and therefore, in combination with the previous analyses, supporting the Double Monotonicity Model. Only 3 significant violations of IIO were observed, involving item 1 with items 4, 6 and 7. The item response functions for item pairs with significant intersections are presented in Figure 3[1]. Because all three violations involved item 1, I computed H_T again without it, and found that the IIO would in this case be strong ($H_T = 0.520$).

[1] For Figure 3, the plotting function of the `mokken` package was modified in order for all rest score groups on the *x*-axis to be consistent. In addition, it can be noted that the three plots involve item 1, but that its item response function appears slightly different in the three plots. The reason for this is that the rest score is computed in each plot using all items but the two items involved in the comparison. Since the item pair is different in each plot, the rest score group is therefore different, leading to slightly different response functions for the same item.

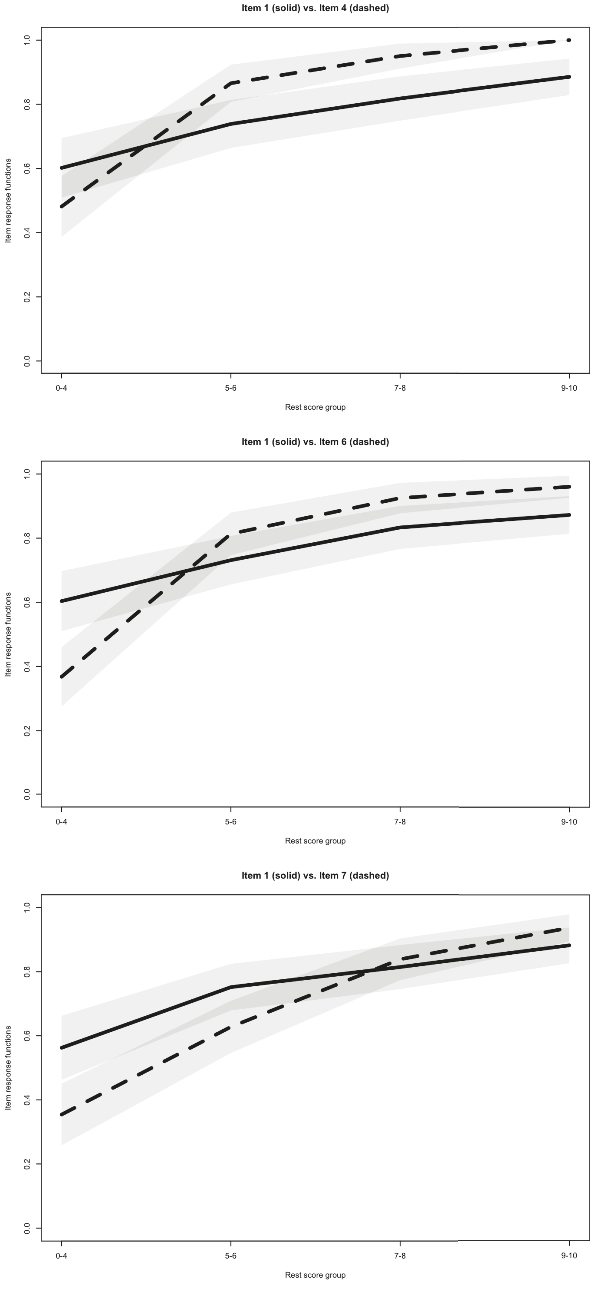


Figure 3. Item response functions (with 95% confidence intervals) of significantly intersecting item pairs.

3.5. Reliability

The MS reliability estimate was 0.836, and the LCRC reliability estimate was 0.876, both indicating, like previously found using other estimates (Myszkowski and Storme 2018), that the SPM-LS had satisfactory reliability. The item-rest correlations ranged between 0.285 and 0.563, item 1 having a notably lower item-rest correlation that the other items. However, the reliability indices were similar without this item ($MS = 0.841$, $LCRC = 0.874$).

4. Discussion

While the SPM-LS has already been investigated using a variety of methods in this very dataset—including parametric IRT, Bayesian IRT, factor analysis, and exploratory graph analysis (Myszkowski and Storme 2018; Garcia-Garzon et al. 2019; Bürkner 2020)—the current study proposes the first investigation of this instrument using non-parametric IRT, and more specifically Mokken Scale Analysis (Mokken 1971; Mokken and Lewis 1982). This framework allowed to study several psychometric properties, permitting to both confirm the previous encouraging results on the SPM-LS—on dimensionality, local independence and reliability—and to investigate new properties—monotonicity and invariant item ordering.

4.1. Conclusions on the SPM-LS

Overall, the SPM-LS showed robust psychometric qualities in this study. More specifically, it was found to have satisfactory monotonicity, scalability, local independence (with only a few local dependencies), invariant item ordering (with only a few significant violations) and reliability. This is an overall satisfactory set of results, which would lead us to encourage the use of this instrument.

The main new elements regarding the investigation of this scale were the support for monotonicity—the item response functions were overall monotonically increasing—and invariant item ordering—the item response functions overall did not intersect, giving, along with unidimensionality and local independence, support for the Double Monotonicity Model. The fact that this model was overall supported is interesting, as it presents several advantages for the use of the SPM-LS in practice (Ligtvoet et al. 2010). First, the monotonicity of item responses suggests that, even though Rasch 1-parameter (and to a lesser extent, 2-parameter) models did not fit well this dataset (Myszkowski and Storme 2018; Bürkner 2020), there is support for the SPM-LS sum scores being able to order persons based on their ability. In addition, it is very clear that each series of Raven's matrices were originally conceptualized as having a cumulative structure, with examinees responding items gradually increasing in difficulty by the stacking of logical rules to decipher and apply: Empirical support for invariant item ordering supports such a hypothetical functioning of the test. Test editors and practitioners generally assume, that, because an item A has a higher success rate than another item B, then item A is necessarily easier than B for all examinees, and they often use a test as though this assumption were true, without empirically testing it (Ligtvoet et al. 2010): The current study provides evidence that it is empirically justified to make such interpretations from the SPM-LS.

It was notable, through this investigation, that the issues encountered tended to involve item 1. More specifically, item 1 was the item with the smallest scalability (based on H_j coefficients), the only one with an outlying negative local dependency (based on W_3 coefficients), was involved in all three significant violations of invariant item ordering, and had the lowest item-rest correlation. While it appears tempting to remove this item, I would recommend to at least maintain it as a training item (meaning, having participants take it but not necessarily including it in the scoring). This is because (1) the presence of this item is still probably important for the examinees to learn the base rule used throughout the series, and (2) the plots suggest that this items' response function is still monotonous, and its intersections with the item response functions of items 4, 6 and 7 appear somewhat minimal, as the confidence intervals overlap for most ability levels. The current study suggests that practitioners and/or future researchers using the SPM-LS use the full instrument, even though they may question and study their own dataset to decide on whether to use item 1 in the scoring or not.

4.2. Limitations

While this investigation presents satisfactory findings regarding the psychometric qualities of the SPM-LS, the different indices observed were not perfect, and notably, the scalability of the scale was only medium (Mokken 1997; Sijtsma and van der Ark 2017), suggesting that the instrument can further be improved. I noted earlier that it would be categorized as strong if item 1 were excluded from the scoring but, albeit strong, it would be still just above the strong threshold. In addition, excluding item 1 from scoring remains a post-hoc suggestion, made after seeing each items' scalability. It would therefore call for further investigations using a new sample.

Mokken Scale Analysis investigates aspects of psychometric instruments that are different from more usual sets of analyses (notably of the factor analytic or Rasch tradition)—especially the investigation of monotonicity and invariant item ordering—but this study also suffers from some limitations of this specific framework. For example, it does not provide a way to study or recover information from distractor responses like other approaches—such as nested logit models (Suh and Bolt 2010), the nominal response model (Bock 1997) or the multiple-choice model (Thissen and Steinberg 1984)—which are an important aspect of this specific test (Myszkowski and Storme 2018). Related to this, MSA certainly allows to graphically study item responses, but, because it is non-parametric, it does not produce item parameters that can be interpreted. This is a limiting factor in this context, because previous results (Myszkowski and Storme 2018; Bürkner 2020) suggest that phenomena like guessing—which is unaccounted for in MSA, apart from potentially appearing in item response functions—are relevant for this test. Another limitation of MSA is that it does not provide a way to investigate conditional reliability, and therefore does not allow to, for example, diagnose if an instrument provides reliable ability estimates across a wide range of ability levels. This is particularly a problem in the case of the SPM-LS, because the fact that it only includes one series of the original SPM implies that the range of abilities that are reliably measured may be limited (Myszkowski and Storme 2018). Finally, other advanced uses of Rasch modeling, such as computer-adaptive testing and test equating, are also impossible with Mokken scaling (Meijer et al. 1990).

4.3. Future Directions

Support for the Double Monotonicity Model, because of invariant item ordering, indicates that, for an item A of lower difficulty than an item B, an examinee who fails item A is predicted to also fail item B (and all items that are more difficult). Thus, if one orders items from the easiest to the most difficult, as is done with the SPM-LS, then it is conceivable to have examinees stop the test after a number of failures. This is because they are likely to then fail all future items. As a supplementary analysis, in this dataset, I computed the correlations between the full scores of examinees (using all item scores) and the scores they would have received, had they been stopped after a number of consecutive failures. I found that stopping the test after only one failure provided scores that were strongly (but far from perfectly) correlated with full scores— $r(483) = 0.735$, $p < 0.001$—while stopping the test after two consecutive items failed would preserve scores nearly perfectly—$r(483) = 0.999$, $p < 0.001$. Based on this, I would suggest that stopping the administration after 2 consecutively failed items could lead to gains of administration time without any substantial loss of information about an examinee's ability. I recommend that future studies further examine this possibility, though the present study already gives quite a strong support for such a use.

Finally, while the psychometric investigation of an instrument can take many shapes, the current study demonstrates how Mokken Scale Analysis can provide insightful information about an instrument, even when that instrument has already been studied in the same dataset with multiple popular and less popular methods (Myszkowski and Storme 2018; Bürkner 2020; Garcia-Garzon et al. 2019.) Besides replicating the present study in other samples and in other conditions—which is certainly called for—I suggest that future studies investigate the SPM-LS using other non-parametric IRT models—for example, spline IRT models (Winsberg et al. 1984)—to better understand its functioning.

Funding: This research received no external funding.

Conflicts of Interest: The author declares no conflict of interest.

References

Bock, R. Darrell. 1997. The Nominal Categories Model. In *Handbook of Modern Item Response Theory*. Edited by Wim J. van der Linden and Ronald K. Hambleton. New York: Springer, pp. 33–49. [CrossRef]

Bors, Douglas A., and Tonya L. Stokes. 1998. Raven's Advanced Progressive Matrices: Norms for First-Year University Students and the Development of a Short Form. *Educational and Psychological Measurement* 58: 382–98. [CrossRef]

Bürkner, Paul-Christian. 2020. Analysing Standard Progressive Matrices (SPM-LS) with Bayesian Item Response Models. *Journal of Intelligence* 8: 5. [CrossRef] [PubMed]

Carpenter, Patricia A., Marcel A. Just, and Peter Shell. 1990. What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review* 97: 404–31. [CrossRef] [PubMed]

Garcia-Garzon, Eduardo, Francisco J. Abad, and Luis E. Garrido. 2019. Searching for G: A New Evaluation of SPM-LS Dimensionality. *Journal of Intelligence* 7: 14. [CrossRef] [PubMed]

Gignac, Gilles 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. [CrossRef]

Hamel, Ronald, and Verena D. Schmittmann. 2006. The 20-Minute Version as a Predictor of the Raven Advanced Progressive Matrices Test. *Educational and Psychological Measurement 66*: 1039–46. [CrossRef]

Junker, Brian W., and Klaas Sijtsma. 2000. Latent and Manifest Monotonicity in Item Response Models. *Applied Psychological Measurement* 24: 65–81. [CrossRef]

Ligtvoet, Rudy, L. Andries Van der Ark, Janneke M. Te Marvelde, and Klaas Sijtsma. 2010. Investigating an Invariant Item Ordering for Polytomously Scored Items. *Educational and Psychological Measurement* 70: 578–95. [CrossRef]

Meijer, Rob R., Klaas Sijtsma, and Nico G. Smid. 1990. Theoretical and Empirical Comparison of the Mokken and the Rasch Approach to IRT. *Applied Psychological Measurement* 14: 283–98. [CrossRef]

Mokkan, Robert J., and Charles Lewis. 1982. A Nonparametric Approach to the Analysis of Dichotomous Item Responses. *Applied Psychological Measurement 6*: 417–30. [CrossRef]

Mokken, Robert Jan. 1971. *A Theory and Procedure of Scale Analysis.* The Hague and Berlin: Mouton de Gruyter.

Mokken, Robert J. 1997. Nonparametric Models for Dichotomous Responses. In *Handbook of Modern Item Response Theory.* Edited by Wim J. van der Linden and Ronald K. Hambleton. New York: Springer, pp. 351–67.[CrossRef]

Molenaar, Ivo W., and Klaas Sijtsma. 2000. *User's Manual MSP5 for Windows*. Groningen: IEC ProGAMMA

Myszkowski, Nils, and Martin Storme. 2018. A snapshot of g? Binary and polytomous item-response theory investigations of the last series of the Standard Progressive Matrices (SPM-LS). *Intelligence* 68: 109–16. [CrossRef]

Rasch, Georg. 1993. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago: MESA Press.

Raven, John C., John Raven, and John Hugh Court. 1962. *Advanced Progressive Matrices*. London: HK Lewis.

Raven, John C. 1941. Standardization of Progressive Matrices, 1938. *British Journal of Medical Psychology* 19: 137–50. [CrossRef]

Ree, Malcolm James, and James A. Earles. 1992. Intelligence Is the Best Predictor of Job Performance. *Current Directions in Psychological Science* 1: 86–89. [CrossRef]

Rohde, Treena Eileen, and Lee Anne Thompson. 2007. Predicting academic achievement with cognitive ability. *Intelligence* 35: 83–92. [CrossRef]

Salgado, Jesús F., Neil Anderson, Silvia Moscoso, Cristina Bertua, Filip De Fruyt, and Jean Pierre Rolland. 2003. A Meta-Analytic Study of General Mental Ability Validity for Different Occupations in the European Community. *The Journal of Applied Psychology* 88: 1068–81. [CrossRef] [PubMed]

Sijtsma, Klaas, and Rob R. Meijer. 1992. A method for investigating the intersection of item response functions in Mokken's nonparametric IRT model. *Applied Psychological Measurement* 16: 149–57. [CrossRef]

Sijtsma, Klaas, and Ivo W. Molenaar. 1987. Reliability of test scores in nonparametric item response theory. *Psychometrika* 52: 79–97. [CrossRef]

Sijtsma, Klaas, and L. Andries van der Ark. 2017. A tutorial on how to do a Mokken scale analysis on your test and questionnaire data. *British Journal of Mathematical and Statistical Psychology* 70: 137–58. [CrossRef]

Sijtsma, Klaas, Rob R. Meijer, and L. Andries van der Ark. 2011. Mokken scale analysis as time goes by: An update for scaling practitioners. *Personality and Individual Differences* 50: 31–37. [CrossRef]

Sijtsma, Klaas. 1998. Methodology Review: Nonparametric IRT Approaches to the Analysis of Dichotomous Item Scores. *Applied Psychological Measurement* 22: 3–31. [CrossRef]

Storme, Martin, Nils Myszkowski, Simon Baron, and David Bernard. 2019. Same Test, Better Scores: Boosting the Reliability of Short Online Intelligence Recruitment Tests with Nested Logit Item Response Theory Models. *Journal of Intelligence* 7: 17. [CrossRef] [PubMed]

Straat, J. Hendrik, L. Andries van der Ark, and Klaas Sijtsma. 2014. Minimum Sample Size Requirements for Mokken Scale Analysis. *Educational and Psychological Measurement* 74: 809–22. [CrossRef]

Straat, J. Hendrik, L. Andries van der Ark, and Klaas Sijtsma. 2016. Using Conditional Association to Identify Locally Independent Item Sets. *Methodology* 12: 117–23. [CrossRef]

Suh, Youngsuk, and Daniel M. Bolt. 2010. Nested Logit Models for Multiple-Choice Item Response Data. *Psychometrika* 75: 454–73. [CrossRef]

Thissen, David, and Lynne Steinberg. 1984. A Response Model for Multiple Choice Items. *Psychometrika* 49: 501–19. [CrossRef]

Tijmstra, Jesper, and Maria Bolsinova. 2019. Bayes Factors for Evaluating Latent Monotonicity in Polytomous Item Response Theory Models. *Psychometrika* 84: 846–69. [CrossRef]

Tijmstra, Jesper, Herbert Hoijtink, and Klaas Sijtsma. 2015. Evaluating Manifest Monotonicity Using Bayes Factors. *Psychometrika* 80: 880–96. [CrossRef]

Van der Ark, L. Andries. 2012. New Developments in Mokken Scale Analysis in R. *Journal of Statistical Software* 48: 1–27. [CrossRef]

Van der Ark, L. Andries. 2007. Mokken Scale Analysis in R. *Journal of Statistical Software* 20: 1–19. [CrossRef]

van der Ark, L. Andries, Daniël W. van der Palm, and Klaas Sijtsma. 2011. A Latent Class Approach to Estimating Test-Score Reliability. *Applied Psychological Measurement* 35: 380–92. [CrossRef]

Winsberg, Suzanne, David Thissen, and Howard Wainer. 1984. Fitting Item Characteristic Curves with Spline Functions. *ETS Research Report Series* 1984. [CrossRef]

Journal of *Intelligence*

MDPI

Article

Regularized Latent Class Analysis for Polytomous Item Responses: An Application to SPM-LS Data

Alexander Robitzsch [1,2]

[1] IPN—Leibniz Institute for Science and Mathematics Education, D-24098 Kiel, Germany; robitzsch@leibniz-ipn.de
[2] Centre for International Student Assessment (ZIB), D-24098 Kiel, Germany

Received: 9 July 2020; Accepted: 10 August 2020; Published: 14 August 2020

Abstract: The last series of Raven's standard progressive matrices (SPM-LS) test was studied with respect to its psychometric properties in a series of recent papers. In this paper, the SPM-LS dataset is analyzed with regularized latent class models (RLCMs). For dichotomous item response data, an alternative estimation approach based on fused regularization for RLCMs is proposed. For polytomous item responses, different alternative fused regularization penalties are presented. The usefulness of the proposed methods is demonstrated in a simulated data illustration and for the SPM-LS dataset. For the SPM-LS dataset, it turned out the regularized latent class model resulted in five partially ordered latent classes. In total, three out of five latent classes are ordered for all items. For the remaining two classes, violations for two and three items were found, respectively, which can be interpreted as a kind of latent differential item functioning.

Keywords: regularized latent class analysis, regularization, fused regularization, fused grouped regularization, distractor analysis

1. Introduction

There has been recent interest in assessing the usefulness of short versions of the Raven's Progressive Matrices. Myszkowski and Storme (2018) composed the last 12 matrices of the Standard Progressive Matrices (SPM-LS) and argued that it could be regarded as a valid indicator of general intelligence g. As part of this special issue, the SPM-LS dataset that was analyzed in Myszkowski and Storme (2018) was reanalyzed in a series of papers applying a wide range of psychometric approaches.

Previous reanalyses of the SPM-LS dataset have in common that quantitative latent variable models were utilized. In this paper, discrete latent variable models (i.e., latent class models) are applied for analyzing the SPM-LS dataset. With discrete latent variable models, the analysis of types instead of traits is the primary focus (see von Davier et al. (2012) and Borsboom et al. (2016)). A disadvantage of discrete latent variable models is that they often have a large number of parameters to estimate. For example, latent class models result in item response probabilities that are allowed to vary across classes. Even with only a few classes, the number of estimated parameters is typically larger than parametric models with quantitative latent variables. Hence, model selection based on principles often favors quantitative latent variable models over discrete latent variable models. So-called regularization approaches automatically reduce the number of parameters to estimate (see Huang et al. (2017) or Jacobucci et al. (2016)) for the use of regularization in structural equation modeling and Tutz and Schauberger (2015) or Battauz (2019) in item response modeling). In this paper, these regularization approaches are applied in discrete latent variable models, and some extensions for polytomous data are proposed.

The paper is structured as follows. In Section 2, we give a brief overview of latent class analysis. In Section 3, regularized latent class analysis for dichotomous and polytomous data is introduced. In Section 4, we apply proposed models of Section 3 in a simulated data illustration. In Section 5, we apply regularized latent class analysis to the SPM-LS dataset. Finally, in Section 6, we conclude with a discussion.

2. Latent Class Analysis

Latent variable models represent discrete items by a number of latent variables (see Agresti and Kateri 2014 for an overview). These latent variables can be categorical or quantitative or a mixture of both. Quantitative latent variables are considered in factor analysis, structural equation models, or item response models. In this article, we focus on categorical latent variables. In this case, latent variables are labeled as latent classes and are extensively studied in the literature of latent class analysis (LCA; Collins and Lanza 2009; Langeheine and Rost 1988; Lazarsfeld and Henry 1968).

A latent class model (LCM) represents the multivariate distribution of I categorical items $\boldsymbol{X} = (X_1, \ldots, X_I)$ by a fixed number of C latent classes. Let U denote the latent class variable that takes one of the values $1, 2, \ldots, C$. It is assumed that items X_i are conditionally independent on the latent class variable U. This means that it holds that

$$P(\boldsymbol{X} = \boldsymbol{x} | U = c) = \prod_{i=1}^{I} P(X_i = x_i | U = c) \quad \text{for } \boldsymbol{x} = (x_1, \ldots, x_I) \quad . \tag{1}$$

The multivariate probability distribution is then given as a mixture distribution

$$P(\boldsymbol{X} = \boldsymbol{x}) = \sum_{c=1}^{K} P(U = c) \prod_{i=1}^{I} P(X_i = x_i | U = c) \quad . \tag{2}$$

Applications of LCMs to intelligence tests can be found in Formann (1982) or Janssen and Geiser (2010).

2.1. Exploratory Latent Class Analysis for Dichotomous Item Responses

In this subsection, we describe the LCM for dichotomous items. Let $p_{ic} = P(X_i = 1 | U = c)$ denote the item response probability for correctly solving item i if a person is located in class c. In the estimation, these bounded parameters ($p_{ic} \in [0, 1]$) are transformed onto the real line by using the logistic transformation (see also Formann 1982)

$$P(X_i = x | U = c) = \frac{\exp(x\gamma_{ic})}{1 + \exp(\gamma_{ic})} \quad (x = 0, 1). \tag{3}$$

Note that p_{ic} is a one-to-one function of γ_{ic}. For estimation purposes, it is sometimes more convenient to estimate models with unbounded parameters instead of estimating models with bounded parameters. For I items and C classes, $I \cdot C$ item parameters have to be estimated in the case of dichotomous items. In comparison to item response models (1PL model: one parameter, 2PL: two parameters, etc.), this results in many more parameters to be estimated. However, LCMs do not pose the assumption that classes are ordered, and no monotonicity assumptions of item response functions are posed.

Moreover, let $p_c = P(U = c)$ denote the probability that a person is in class c. As for item parameters, a logistic transformation is used to represent the class probabilities p_c by parameters δ_c. More formally, we set

$$p_c = \frac{\exp(\delta_c)}{1 + \sum_{j=2}^{C} \exp(\delta_j)} \quad (c = 1, \ldots, C), \tag{4}$$

where $\delta_1 = 0$. Because the probabilities sum to one, only $C - 1$ distribution parameters have to be estimated. In total, the saturated distribution of I dichotomous items has $2^I - 1$ free possible parameters, which is represented by $I \cdot C + C - 1$ parameters in the LCM with C classes.

LCMs can be interpreted as pure exploratory models because no structure of item response probabilities among classes is posed. Confirmatory LCMs assume additional equality constraints on item response probabilities (Finch and Bronk 2011; Nussbeck and Eid 2015; Oberski et al. 2015; Schmiege et al. 2018). Like in confirmatory factor analysis, it could be assumed that some items load only on some classes, which translates into equal item response probabilities. Cognitive diagnostic models can be seen as particular confirmatory LCMs (von Davier and Lee 2019).

It should be emphasized that restricted LCMs form the basis of almost all popular latent variable models for discrete item responses that are nowadays very popular. Formann (1982) suggested to represent the vector $\boldsymbol{\gamma} = (\gamma_{11}, \ldots, \gamma_{1C}, \ldots, \gamma_{I1}, \ldots \gamma_{IC})$ of item parameters as linear combinations $\gamma_{ic} = \boldsymbol{q}_{ic}\boldsymbol{\alpha}$ $(i = 1, \ldots, I; c = 1, \ldots, C)$ using a parameter vector $\boldsymbol{\alpha}$ and known weight vectors $\boldsymbol{q}_{ic}$. In addition, the distribution parameter $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_C)$ is represented by $\delta_c = \boldsymbol{w}_c\boldsymbol{\beta}$ using a parameter vector $\boldsymbol{\beta}$ and known weight vectors $\boldsymbol{w}_c$. The resulting so-called structured LCM (Formann and Kohlmann 1998) includes unidimensional and multidimensional 1PL and 2PL models as special cases as well as mixture item response models (Formann 2007). To accomplish this, continuous latent variables are approximated by a finite number of discrete latent classes. For example, a normally distributed latent variable is approximated by discrete latent classes (e.g., $C = 21$ classes) whose probabilities are represented by only two components in α (i.e., the mean and the standard deviation as the first two moments). The usage of discrete latent classes can be interpreted as performing numerical integration with a fixed integration grid and applying the rectangle rule. Similar generalizations of restricted LCM were proposed by researcher von Davier (2008, 2010). In the rest of this article, we will focus on the simple exploratory LCM, although the proposed extension also applies to the more general structured latent class model.

In many applications, the allocation of persons to classes should be predicted by person variables $\boldsymbol{Z}$ (Collins and Lanza 2009). In more detail, class probabilities $p_c = P(U = c)$ are replaced by subject-specific conditional probabilities $P(U = c|\boldsymbol{Z} = \boldsymbol{z})$ (so-called latent class regression). These models further ease the interpretation of latent classes.

2.2. Exploratory Latent Class Analysis for Polytomous Item Responses

Now assume that there are I polytomous items and each item has $K_i + 1$ nominal categories $0, 1, \ldots, K_i$. Item response probabilities are then given as $p_{ikc} = P(X_i = k|U = c)$ and are again transformed into unbounded parameters γ_{ikc} by a logistic transformation. In more detail, it is assumed that

$$P(X_i = x|U = c) = \frac{\exp(\gamma_{ixc})}{1 + \sum_{k=1}^{K_i} \exp(\gamma_{ikc})} \quad (x = 0, 1, \ldots, K_i), \tag{5}$$

where $\gamma_{i0c} = 0$ for all items i and all latent classes c. Instead of estimating $K_i + 1$ probabilities for item i and class c, K_i free parameters γ_{ihc} have to be estimated. If all polytomous items have $K + 1$ categories, the multidimensional contingency table of observations has $(K + 1)^I - 1$ free parameters while in the LCM $I \cdot K \cdot C + C - 1$ parameters are estimated. It should be emphasized that the LCM for polytomous items has more free parameters compared to LCMs with dichotomous items as well as for unidimensional and multidimensional item response models for polytomous data.

Like for dichotomous data, restricted LCMs were formulated that represented the vector of all item response functions by $\gamma_{ikc} = \boldsymbol{q}_{ikc}\boldsymbol{\alpha}$ $(i = 1, \ldots, I; k = 1, \ldots, K_i; c = 1, \ldots, C)$ using a parameter vector $\boldsymbol{\alpha}$ and known weight vectors $\boldsymbol{q}_{ikc}$ (Formann 1992). It is advisable to induce some structure on item response functions, especially for polytomous data, because many parameters have to be estimated without any structural assumptions.

3. Regularized Latent Class Analysis

As LCMs are exploratory models, interpretation of results could sometimes be challenging. Moreover, in not too large samples, parameter estimation gets instable, and findings are sometimes not generalizable across different samples. Alternatively, confirmatory latent models could be estimated for obtaining more stable and more interpretable parameter estimates. However, such confirmatory approaches need assumptions that have to be known in advance of the data analysis. Hence, alternative approaches are sought.

Regularized latent class models (RLCMs; Chen et al. 2017) estimate item response probabilities under the presupposition that similar item response probabilities in these models are grouped and receive the same value. The main idea of using the regularization technique (see Hastie et al. 2015 for an overview) to LCMs is that by subtracting an appropriate penalty term from the log-likelihood function, some simpler structure on item response probabilities is posed. Different penalty terms typically result in different estimated parameter structures. In a recent Psychometrika paper, Chen et al. (2017) proposed the RLCM for dichotomous item responses. Related work for dichotomous data can be found in Wu (2013) and Yamamoto and Hayashi (2015).

The regularization technique has also been applied for factor models with continuous items (Huang et al. 2017; Jacobucci et al. 2016) and discrete items (Chen et al. 2018; Sun et al. 2016) in order to fit exploratory factor models with the goal of estimating as many zero loadings as possible. In this respect, regularization is a viable alternative to factor rotation methods (Scharf and Nestler 2019).

The regularization technique has also been applied to Gaussian mixture models in which cluster means are estimated to be equal for some variables among clusters (Bhattacharya and McNicholas 2014; Ruan et al. 2011). Regularized latent class analysis (RLCA) is also referred to as penalized latent class analysis (see DeSantis et al. 2008). Under this label, LCMs are typically meant by that apply regularization to the estimation of regression coefficients of the latent class regression model (DeSantis et al. 2008, 2012; Houseman et al. 2006; Leoutsakos et al. 2011; Sun et al. 2019; Wu et al. 2018). Fop and Murphy (2018) provide a recent review of applications of the regularization technique in mixture models.

In the following, we describe the RLCM at first for dichotomous items. Afterward, we consider the more complex case of polytomous items in which more possibilities for setting equality constraints among item response probabilities are present.

3.1. Regularized Latent Class Analysis for Dichotomous Item Responses

At first, we consider the case of dichotomous items X_i $(i = 1, \ldots, I)$. In an RLCM, not all item response probabilities p_{ic} $(c = 1, \ldots, C)$ are assumed to be unique. Chen et al. (2017) subtracted a penalty term from the log-likelihood function that penalizes differences in ordered item response probabilities. In more detail, denote by $p_{i,(c)}$ $(c = 1, \ldots, C)$ ordered item response probabilities of the original probabilities p_{ic}

such that $p_{i,(1)} \leq p_{i,(2)} \leq \ldots p_{i,(C)}$, and collect all parameters in $\boldsymbol{p}_i^*$. Then, Chen and colleagues used the following penalty function for item i

$$Pen(\boldsymbol{p}_i^*; \lambda) = \sum_{c=2}^{C} H_{\text{SCAD}}(p_{i,(c)} - p_{i,(c-1)}; \lambda), \tag{6}$$

where H_{SCAD} denotes the smoothly clipped absolute deviation penalty (SCAD; Fan and Li 2001). The SCAD penalty takes a value of zero if $p_{i,(c)} - p_{i,(c-1)} = 0$ and is positive otherwise (see Figure 1 for the functional form of the SCAD penalty). The parameter λ is a regularization parameter that governs the strength of the penalty function. With small values of λ, differences are barely penalized, but with large values of λ, differences are heavily penalized, and item parameters approach a uniform distribution.

If $\boldsymbol{X}$ denotes the matrix of observed data and $\boldsymbol{p}^*$ denotes the vector of all ordered item response probability and $\boldsymbol{\delta}$ the vector that represents the skill class probabilities, the following function is maximized in Chen et al. (2017):

$$l(\boldsymbol{p}^*, \boldsymbol{\delta}; \boldsymbol{X}) - N \sum_{i=1}^{I} Pen(\boldsymbol{p}_i^*; \lambda), \tag{7}$$

where l denotes the log-likelihood function of the data. By employing a penalty function Pen in the estimation, some item response probabilities are merged, which, in turn, eases the interpretation of resulting latent classes. It should be noted that for estimating model parameters, the regularization parameter λ has to be fixed. In practice, the regularization parameter λ also has to be estimated. Hence, the maximization is performed on a grid of λ values (say, $\lambda = 0.01, 0.02, \ldots, 0.30$), and that model is selected that is optimal with respect to some criterion. Typical criteria are the cross-validated log-likelihood or information criteria like the Akaike information criterion (AIC), Bayesian information criterion (BIC), or others (Hastie et al. 2015).

The maximization of (7) is conducted using an expectation-maximization (EM) algorithm (see Section 3.3 for general description). The estimation approach of Chen et al. (2017) is implemented in the R package CDM (George et al. 2016; Robitzsch and George 2019).

3.1.1. Fused Regularization among Latent Classes

Though the estimation approach of Chen et al. (2017) is successful, it is not clear how it could be generalized to polytomous data because it is not evident how item response probabilities of several categories should be ordered. Hence, we propose a different estimation approach. We apply the technique of fused regularization (Tibshirani et al. 2005; Tutz and Gertheiss 2016) that penalizes all pairwise differences of item response probabilities. In more detail, for a vector $\boldsymbol{p}_i$ of item response probabilities, we replace the penalty (used in Equation (6)) of Chen et al. (2017) by

$$Pen(\boldsymbol{p}_i; \lambda) = \sum_{c<d} H_{\text{MCP}}(p_{ic} - p_{id}; \lambda), \tag{8}$$

where $p_{ic} = P(X_i = 1|U = c)$ are class-specific item response probabilities, and h_{MCP} denotes the minimax concave penalty (MCP; Zhang 2010). We do not suppose dramatic differences to the SCAD penalty, but we would expect less biased estimators than using the often employed least absolute shrinkage and selection operator (LASSO) penalty $H_{\text{LASSO}}(x; \lambda) = \lambda |x|$ (see Hastie et al. 2015). By using pairwise differences in Equation (8), item response probabilities are essentially merged into item-specific clusters of values that are equal within each cluster. Hence, the same goal as in Chen et al. (2017) is achieved. As explained in

Section 2.2, our estimation approach uses transformed item response probabilities γ. Therefore, in the estimation, we replace Equation (8) by

$$Pen(\gamma_i; \lambda) = \sum_{c<d} H_{\text{MCP}}(\gamma_{ic} - \gamma_{id}; \lambda). \quad (9)$$

Note that by using the penalty on γ_i in Equation (9) instead of on $\boldsymbol{p}_i$ in Equation (8), a different metric in quantifying differences in item parameters is introduced. By using γ_i, differences in extreme probabilities (i.e., probabilities near 0 or 1) appear to be less similar than by using untransformed probabilities as in (8).

In Figure 1, the LASSO, MCP, and SCAD penalty functions are depicted. It can be seen for x values near to 0, the MCP and the SCAD penalty equal the LASSO penalty (i.e., $f(x) = \lambda|x|$). For sufficiently large x values MCP and SCAD reach an upper asymptote, which is not the case for the LASSO penalty. Hence, for the MCP and SCAD penalty, the penalty is relatively constant for large values of x. This property explains why the MCP and SCAD penalty typically results in less biased estimates. It should be noted that the application of the regularization presupposes some sparse structure in the data for obtaining unbiased estimates. In other words, the true data generating mechanism consists of a sufficiently large number of equal item parameters. If all item response probabilities would be different in the data generating model, employing a penalty that forces many item parameters to be equal to each other would conflict the data generating model.

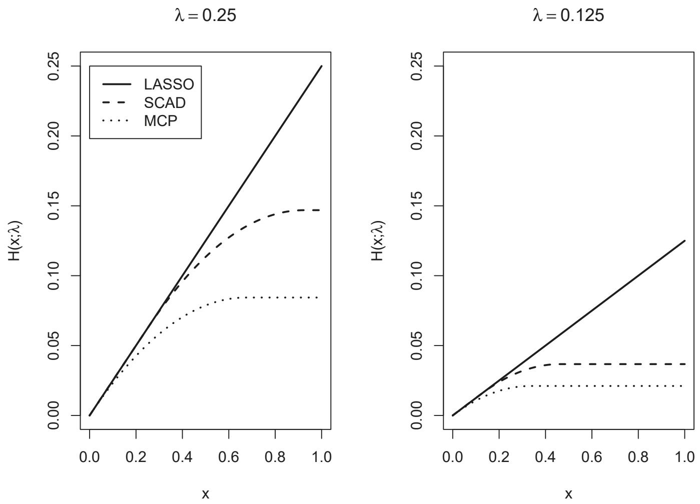


Figure 1. Different penalty functions used in regularization with regularization parameter $\lambda = 0.25$ (**left panel**) and $\lambda = 0.125$ (**right panel**).

3.1.2. Hierarchies in Latent Class Models

The RLCM can be used to derive a hierarchy among latent classes. The main idea is depicted in Figure 2. In an RLCM with $C = 4$ classes, a partial order of latent classes is defined. Class 1 is smaller than Classes 2, 3, and 4. Classes 2 and 3 cannot be ordered. Finally, Classes 2 and 3 are smaller than Class 4. More formally, in an RLCM, we define class c to be *smaller* than class d (or: class d is *larger* than class c) if all item response probabilities in class c are at most as large as in class d, i.e., $p_{ic} \leq p_{id}$ for all items $i = 1, \ldots, I$.

We use the notation $c \preceq d$ to indicate that c is smaller than d. In a test with many items, fulfilling these inequalities for all items might be a too strong requirement. Hence, one weakens the concept of partial ordering a bit. Given a tolerable for at most ι items, we say that class c is *approximately smaller* than class d if $p_{ic} \leq p_{id}$ is fulfilled for at least $I - \iota$ items.

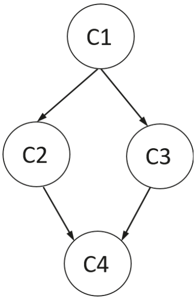


Figure 2. Illustration of a partial order with four latent classes.

The partial ordering of latent classes substantially eases the interpretation of the results in RLCMs. Chen et al. (2017) used the RLCM to derive partially ordered latent classes in cognitive diagnostic modeling. Wang and Lu (2020) also applied the RLCM for estimating hierarchies among latent classes (see also Robitzsch and George 2019). Using the RLCM with an analysis of hierarchies may be considered as a preceding method of confirmatory approaches to latent class modeling.

3.2. Regularized Latent Class Analysis for Polytomous Item Responses

In the following, we propose an extension of RLCM for polytomous item responses. It has been shown that using information from item distractors (Myszkowski and Storme 2018; Storme et al. 2019) could increase the reliability for person ability estimates compared to using only dichotomous item responses that only distinguishes between correct and incorrect item responses. Moreover, it could be beneficial to learn about the differential behavior of item distractors analyzing the data based on correct and all incorrect item responses.

Assume that 0 denotes the category that refers to a correct response and $1, \ldots, K_i$ refer to the categories of the distractors. In our parameterization of the LCM for polytomous data (see Section 2.2), only parameters γ_{ikc} of distractors k for item i in classes c are parameterized. Given the relatively small sample size of the SPM-LS application data (i.e., $N = 499$), the number of estimated parameters in an unrestricted LCM turn out to be quite large because there are seven distractors per item. Moreover, it could be supposed that the distractors of an item behave similarly. Hence, it would make sense to estimate some item parameters to be equal to each other.

We now outline alternatives for structural assumptions on item response probabilities. Let us fix item i. For $K_i + 1$ categories and C classes, $K_i \cdot C$ item parameters are modeling (omitting the category 0). Hence, we can distinguish between different strategies to the setting of equalities of item parameters. First, for a fixed category k, one can merge some item response probabilities among classes. This means that some of the differences $\gamma_{ikc} - \gamma_{ikd}$ ($c \neq d$) are zero. Hence, a penalty on differences $\gamma_{ikc} - \gamma_{ikd}$ has to be posed. This is just the penalty as for dichotomous items (see Equation (8)), but the regularization is applied for K_i categories instead of one category. Second, for a fixed class c, some item response probabilities among categories could be merged. In this case, one would impose a penalty on differences $\gamma_{ikc} - \gamma_{ihc}$ ($k \neq h$).

Third, penalization among classes and among categories can be simultaneously applied. In the remainder, we outline the different strategies in more detail.

3.2.1. Fused Regularization among Latent Classes

Let $\gamma_{ik*} = (\gamma_{ik1}, \ldots, \gamma_{ikC})$ denote the vector of item parameters for item i in category k. Again, let γ_i denote the vector of all item parameters of item i. For a regularization parameter λ_1 and item i, we define the penalty

$$Pen(\gamma_i; \lambda_1) = \sum_{k=1}^{K_i} Pen(\gamma_{ik*}; \lambda_1) = \sum_{k=1}^{K_i} \sum_{c<d} H_{\mathrm{MCP}}(\gamma_{ikc} - \gamma_{ikd}; \lambda_1). \tag{10}$$

As a result, for a category, some item response probabilities will be merged across latent classes. However, the merging of item parameters (also referred to as fusing; Tibshirani et al. 2005) is independently applied for all categories of an item. In practice, it is maybe not plausible that all distractors of an item would function differently, and item parameters should be more regularized.

3.2.2. Fused Regularization among Categories

As a second alternative, we now merge categories. Let $\gamma_{i*c} = (\gamma_{i1c}, \ldots, \gamma_{iK_ic})$ denote the vector of item parameters for item i in class c. For a regularization parameter λ_2 and item i, we define the penalty

$$Pen(\gamma_i; \lambda_2) = \sum_{c=1}^{C} Pen(\gamma_{i*c}; \lambda_2) = \sum_{c=1}^{C} \sum_{k<h} H_{\mathrm{MCP}}(\gamma_{ikc} - \gamma_{ihc}; \lambda_2). \tag{11}$$

As a result, some of the item response probabilities of categories are set equal to each other. As an outcome of applying this penalty, atypical distractors could be detected. However, by using the penalty in Equation (11), no equalities among latent classes are imposed.

3.2.3. Fused Regularization among Latent Classes and Categories

The apparent idea is to combine the regularization among latent classes and categories. By doing so, the penalties in Equations (10) and (11) have to be added. In more detail, for regularization parameters λ_1 and λ_2, we use the penalty

$$Pen(\gamma_i; \lambda_1, \lambda_2) = \sum_{k=1}^{K_i} \sum_{c<d} H_{\mathrm{MCP}}(\gamma_{ikc} - \gamma_{ikd}; \lambda_1) + \sum_{c=1}^{C} \sum_{k<h} H_{\mathrm{MCP}}(\gamma_{ikc} - \gamma_{ihc}; \lambda_2). \tag{12}$$

It can be seen that the penalty in Equation (12) now depends on two regularization parameters. In the estimation, the one-dimensional grid of regularization parameters has then to be substituted by a two-dimensional grid. This substantially increases the computational demand.

3.2.4. Fused Group Regularization among Categories

We can now proceed to pose additional structural assumptions on item parameters. One could suppose that two distractors k and h of item i show the same behavior. In the RLCM, this means that $\gamma_{ikc} - \gamma_{ihc} = 0$ holds for all classes $c = 1, \ldots, C$. The group regularization technique allows us to estimate all parameters in a subset of parameters to be zero (see Huang et al. 2012 for a review). A fused group regularization approach presupposes that either all differences $\gamma_{ikc} - \gamma_{ihc}$ equal zero or all differences are estimated to be different from zero (Cao et al. 2018; Liu et al. 2019). This property can be achieved by substituting a norm of the difference of the two vectors in the penalty. In more detail, one considers the penalty

$$Pen(\gamma_i;\lambda_1) = \sum_{k<h} H_{\mathrm{MCP}}(||\gamma_{ik*} - \gamma_{ih*}||;\lambda_1) \tag{13}$$

where for a vector $\boldsymbol{x} = (x_1, \ldots, x_p)$, the norm $||\boldsymbol{x}||$ is defined as $||\boldsymbol{x}|| = \sqrt{p}\sqrt{\sum_{k=1}^{p} x_k^2}$. In practice, using the penalty in Equation (13) could provide a more parsimonious estimation than the penalty defined in Equation (12). In principle, model comparisons can be carried out to decide which assumption is better represented in the data.

3.2.5. Fused Group Regularization among Classes

Alternatively, one could also assume that latent classes function the same among classes. In the RLCM, then it would hold that that $\gamma_{ikc} - \gamma_{ikd} = 0$ for all categories $k = 1, \ldots, K_i$. A fused group regularization results in the property that either all item parameters of classes c and d are equal to each other or all estimated to be different from each other. The following penalty is used in this case:

$$Pen(\gamma_i;\lambda_2) = \sum_{c<d} H_{\mathrm{MCP}}(||\gamma_{i*c} - \gamma_{i*d}||;\lambda_2) \tag{14}$$

3.3. Estimation

We now describe the estimation of the proposed RLCM for polytomous data. Let $\boldsymbol{X} = (x_{nik})$ denote the observed dataset where x_{nik} equals 1 if person n ($n = 1, \ldots, N$) chooses category k for item i. Let γ_i denote item parameters of item i and the vector that contains item parameters of all items. The vector $\boldsymbol{\delta}$ represents the skill class distribution. Furthermore, let $p_{ic}(x;\gamma_i) = P(X_i = x|U = c)$ and $p_c(\boldsymbol{\delta}) = P(U = c)$.

Following Chen et al. (2017) and Sun et al. (2016), an EM algorithm is applied for estimating model parameters. The complete-data log-likelihood function is given

$$l_{\mathrm{com}}(\gamma, \delta, \boldsymbol{U}) = \sum_{n=1}^{N}\sum_{i=1}^{I}\sum_{k=1}^{K_i}\sum_{c=1}^{C} x_{nik}u_{nc}\log p_{ic}(k;\gamma_i) + \sum_{n=1}^{N}\sum_{c=1}^{C} u_{nc}\log p_c(\boldsymbol{\delta}), \tag{15}$$

where $\boldsymbol{u}_n = (u_{n1}, \ldots, u_{nC})$ is the vector of latent class indicators for person n. It holds that $u_{nc} = 1$ if person n is located in class c. Obviously, the true class membership is unknown and, hence, Equation (15) cannot be used for maximization.

In the EM algorithm, the estimation of l_{com} is replaced by the expected complete-data log-likelihood function by integrating over the posterior distribution. In more detail, unobserved values u_{nc} are replaced by their conditional expectations:

$$u_{nc}^* = E(u_{nc}|\boldsymbol{x}_n;\boldsymbol{\gamma}^{(t)},\boldsymbol{\delta}^{(t)}) = \frac{p_c(\boldsymbol{\delta}^{(t)})\prod_{i=1}^{I}\prod_{k=1}^{K_i} p_{ic}(k;\boldsymbol{\gamma}_i^{(t)})^{x_{nki}}}{\sum_{d=1}^{D} p_d(\boldsymbol{\delta}^{(t)})\prod_{i=1}^{I}\prod_{k=1}^{K_i} p_{id}(k;\boldsymbol{\gamma}_i^{(t)})^{x_{nki}}} \quad (c = 1, \ldots, C), \tag{16}$$

where $\gamma^{(t)}$ and $\delta^{(t)}$ are parameter estimates from a previous iteration t. The EM algorithm alternates between the E-step and the M-step. By replacing the unobserved values u_{ni} by their expected values u_{ni}^*, the following Q-function is obtained that is used for maximization in the M-step

$$Q(\gamma, \delta | \gamma^{(t)}, \delta^{(t)}) = \sum_{n=1}^{N} \sum_{i=1}^{I} \sum_{k=1}^{K_i} \sum_{c=1}^{C} x_{nik} u^*_{nc} \log p_{ic}(k; \gamma_i) + \sum_{n=1}^{N} \sum_{c=1}^{C} u^*_{nc} \log p_c(\delta). \tag{17}$$

From this Q-function, the penalty function is subtracted such that the following function is minimized for some regularization parameter λ in the M-step

$$Q(\gamma, \delta | \gamma^{(t)}, \delta^{(t)}) - N \sum_{i=1}^{I} Pen(\gamma_i; \lambda). \tag{18}$$

It can be seen that item parameters γ_i are separately obtained for each item i in the M-step because the penalties are defined independently for each item. Hence, for each item i, one maximizes

$$\sum_{n=1}^{N} \sum_{k=1}^{K_i} \sum_{c=1}^{C} x_{nik} u^*_{nc} \log p_{ic}(k; \gamma_i) - NPen(\gamma_i; \lambda). \tag{19}$$

Latent class probability parameters δ are also obtained independently from item parameters in the M-step.

The penalty function *Pen* turns out to be non-differentiable. Here, we use a differentiable approximation of the penalty function (Oelker and Tutz 2017; see also Battauz 2019). As it is well known that the log-likelihood function in LCMs is prone to multiple maxima, using multiple starting values in the estimation is advised.

The described EM algorithm is included in an experimental version of the function `regpolca()` in the R package `sirt` (Robitzsch 2020). The function is under current development for improving computational efficiency.

4. Simulated Data Illustration

Before we illustrate the application of the method to the SPM-LS dataset, we demonstrate the technique using a simulated data set. This helps to better understand the proposed method of regularized latent class modeling under ideal conditions.

4.1. Dichotomous Item Responses

4.1.1. Data Generation

First, we consider the case of dichotomous items. To mimic the situation in the SPM-LS dataset, we also chose $I = 12$ items for simulating a dataset. Moreover, to reduce sampling uncertainty somewhat, a sample size of $N = 1000$ subjects was chosen. There were $C = 4$ latent classes with true class probabilities 0.30, 0.20, 0.10, and 0.40. In Table 1, we present the item response probabilities with each cluster. We only specified parameters for six items and duplicated these parameters for the remaining six items in the test. It can be seen in Table 1 that many item response probabilities were set equal to each other. Indeed, for the first four items, there are only two instead of four unique probabilities. Moreover, it is evident from Table 1 that the four classes are partially ordered. The first class has the lowest probabilities for all items and is, therefore, the smallest class that consists of the least proficient subjects. The fourth class has the highest probabilities, constitutes the largest class, and contains the most proficient subjects.

The model selection is carried out using information criteria AIC and BIC. For regularized models, the required number of parameters in the computation of information criteria is determined by the number of estimated unique parameters. For example, if four item response probabilities would be estimated to be equal in a model, only one parameter would be counted.

Table 1. Data illustration dichotomous data: true item response probabilities p_{ic}.

Item	Class			
	1	**2**	**3**	**4**
1, 7	0.10	0.82	0.82	0.82
2, 8	0.22	0.88	0.88	0.88
3, 9	0.16	0.79	0.16	0.79
4, 10	0.25	0.85	0.25	0.85
5, 11	0.10	0.10	0.46	0.91
6, 12	0.22	0.22	0.22	0.79

4.1.2. Results

In the first step, we estimated exploratory latent class models with $C = 2, 3, 4, 5$, and 6 classes. The model comparison is presented in Table 2. While the decision based on the AIC was ambiguous and selected the incorrect number of classes, the BIC correctly selected model with $C = 4$ latent classes. This observation is consistent with the literature that argues that model selection in LCMs should be based on the BIC instead of the AIC (Collins and Lanza 2009; Keribin 2000).

Table 2. Data illustration dichotomous data: model comparison for exploratory latent class models (LCMs).

C	#np	AIC	BIC
2	25	13,636	13,759
3	38	13,169	13,356
4	51	12,979	**13,229**
5	64	12,981	13,295
6	76	**12,976**	13,349

Note: C = number of classes; #np = number of estimated parameters.

In the solution with four classes, estimated class probabilities were 0.290, 0.204, 0.120, and 0.386, respectively, which closely resembled the data generating values. In Table 3, estimated item response probabilities are shown. The estimates were very similar to the data generating parameters that are presented in Table 1. It can be seen that some deviations from the simulated equality constraints are obtained. It is important to emphasize that latent class solutions are not invariant with respect to their class labels (so-called label switching). Class labels in the estimated model have to be permuted in order to match the class label in the simulated data.

Finally, we estimated the RLCM for regularization parameters from 0.01 to 1.00 in steps of 0.01 for $C = 4$ classes in order to obtain the best-fitting solution. The regularization parameter $\lambda = 0.21$ provided the best-fitting model in terms of the BIC (BIC $= 13{,}104$, AIC $= 12{,}957$). Notably, this model showed a substantially better model fit than the exploratory LCM with four classes due to the more parsimonious estimation of item parameters. In the model with $\lambda = 0.21$, in total, 21 item parameters were regularized (i.e., they were set equal to item parameters in other classes for the respective item), resulting in 30 freely estimated model parameters. With respect to the AIC, the best-fitting model was obtained for $\lambda = 0.15$ (BIC $= 13{,}110$, AIC $= 12{,}953$). This model resulted in 19 regularized item parameters and 32 freely estimated item parameters. The model selected by the minimal BIC ($\lambda = 0.21$) resulted in estimated class probabilities of 0.29, 0.21, 0.11, and 0.39. Estimated item response probabilities (shown in Table 4) demonstrate that the equality constraints that were posed in the data generating were correctly identified in the estimated model.

Table 3. Data illustration dichotomous data: estimated item response probabilities in exploratory LCM with $C = 4$ classes.

Item	Class			
	1	**2**	**3**	**4**
1	0.08	0.79	0.79	0.82
2	0.20	0.84	0.89	0.91
3	0.15	0.76	0.19	0.81
4	0.27	0.90	0.29	0.86
5	0.10	0.09	0.44	0.92
6	0.23	0.23	0.30	0.77
7	0.07	0.79	0.82	0.85
8	0.24	0.87	0.91	0.87
9	0.14	0.82	0.18	0.81
10	0.19	0.90	0.42	0.83
11	0.10	0.13	0.54	0.89
12	0.25	0.19	0.19	0.77

Table 4. Data illustration dichotomous data: estimated item response probabilities in the regularized latent class model (RLCM) with $C = 4$ classes based on the minimal Bayesian information criterion (BIC) ($\lambda = 0.21$).

Item	Class			
	1	**2**	**3**	**4**
1	0.08	0.80	0.80	0.80
2	0.20	0.88	0.88	0.88
3	0.15	0.79	0.15	0.79
4	0.27	0.87	0.27	0.87
5	0.09	0.09	0.44	0.92
6	0.23	0.23	0.23	0.76
7	0.07	0.82	0.82	0.82
8	0.24	0.87	0.87	0.87
9	0.14	0.81	0.14	0.81
10	0.19	0.85	0.40	0.85
11	0.11	0.11	0.55	0.89
12	0.22	0.22	0.22	0.76

Lastly, we want to illustrate the behavior of regularization. For the sequence of specified regularization parameters λ in the estimation, the estimated item response probabilities $p_{ic}(\lambda)$ can be plotted. Such a plot is also referred to as a regularization path (Hastie et al. 2015). With very small λ values, no classes were merged, and all item parameters were estimated differently from each other. With increasing values of λ, item parameters were subsequently merged. For Item 1 and Classes 2, 3, and 4, the regularization path is shown in Figure 3. At first, item parameters of Class 2 and 4 are merged at $\lambda = 0.04$. Afterwards, all three item response probabilities are merged at $\lambda = 0.09$.

4.2. *Polytomous Item Responses*

4.2.1. Data Generation

Now, we simulate illustrated data with polytomous item responses with 12 items, each item possessing four categories (i.e., $K_i = 3$). The first category (i.e., Category 0) refers to the correct category, while categories 1, 2, and 3 refer to distractors of the item. As in the case of dichotomous data, item response probabilities were 0.30, 0.20, 0.10, and 0.40, and $N = 1000$ subjects were simulated. Again, we specified item parameters for the first six items and replicated the parameters for the remaining six items. In Table 5, true item response probabilities are shown that were used for generating the dataset. It is evident that the item response probabilities are strongly structured. All distractors of Items 1 and 5 function precisely the same. For Item 2, Category 1, and Category 2 show the same behavior. Category 3 only shows a differential behavior in Classes 2 and 4. At the other extreme, all item response probabilities differ for Item 6 among classes and categories. It can be expected that an RLCM will result in a substantial model improvement compared to an exploratory LCM without equality constraints.

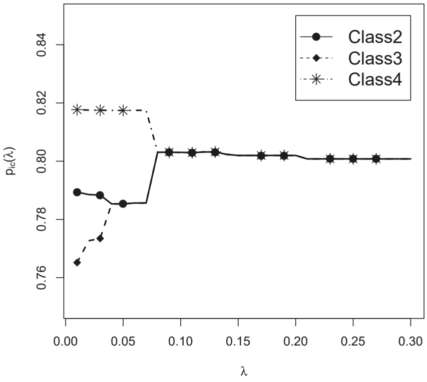


Figure 3. Data illustration dichotomous data: Regularization path for estimated item response probabilities for Item 1 for Classes 2, 3, 4 for the four-class solution.

4.2.2. Results

At first, we fitted exploratory LCMs with 2, 3, 4, 5, and 6 classes. Based on the information criteria presented in Table 6, the correct model with $C = 4$ latent classes was selected. However, the difference in model improvement by moving from 3 to 4 classes would be considered as negligible (i.e., a BIC difference of 3) in practice. Estimated latent class probabilities in the model with four latent classes were estimated as 0.29, 0.21, 0.12, and 0.38. Estimated item response probabilities are shown in Table A1 in Appendix A.

In the next step, different RLCMs for polytomous data were specified. As explained in Section 3.2, one can regularize differences in item parameters among classes (using a regularization parameter λ_1), among categories (using a regularization parameter λ_2), or both (using both regularization parameters or applying fused group regularization). We fitted the five approaches (Approaches R1, …, R5) that were introduced in Section 3.2 to the simulated data using unidimensional and two-dimensional grids of regularization parameters. For each of the regularization approaches, we selected the model with minimal BIC.

In Table 7, it can be seen that the model with the fused penalty on item categories fitted the model best (Approach R2: BIC = 24,689). In this model, 79 item parameters are regularized. The decrease in BIC compared to an exploratory LCM is substantial. From these findings, it follows for this dataset that it is important to fuse item parameters among categories instead of among classes. The best-fitting model when using a simultaneous penalty for classes and categories (Approach R3: BIC = 24,836) outperformed the model in which only parameters were fused among classes (Approach R1: BIC = 24,932). However, it was inferior to the model with fusing among categories. Notably, the largest number of regularized parameters (#nreg = 103) was obtained for Approach R3. The fused grouped regularization approaches (Approaches R4 and R5) also improved fit compared to an unrestricted exploratory LCM but were also inferior to R2. The reason might be that applying group regularization results in the extreme decision that either all item parameters are equal or all are different. In contrast, fused regularization Approaches R1, R2, and R3 allow the situation in which only some of the item parameters are estimated to be equal to each other.

Table 5. Data illustration polytomous data: true item response probabilities p_{ic}.

Item	Cat	Class				Item	Cat	Class			
		1	2	3	4			1	2	3	4
1, 7	0	0.10	0.82	0.82	0.82	4, 10	0	0.25	0.85	0.25	0.85
1, 7	1	0.30	0.06	0.06	0.06	4, 10	1	0.35	0.03	0.35	0.03
1, 7	2	0.30	0.06	0.06	0.06	4, 10	2	0.20	0.03	0.20	0.03
1, 7	3	0.30	0.06	0.06	0.06	4, 10	3	0.20	0.09	0.20	0.09
2, 8	0	0.22	0.88	0.88	0.88	5, 11	0	0.10	0.10	0.46	0.91
2, 8	1	0.26	0.05	0.04	0.06	5, 11	1	0.30	0.30	0.18	0.03
2, 8	2	0.26	0.05	0.04	0.06	5, 11	2	0.30	0.30	0.18	0.03
2, 8	3	0.26	0.02	0.04	0.00	5, 11	3	0.30	0.30	0.18	0.03
3, 9	0	0.16	0.79	0.16	0.79	6, 12	0	0.22	0.22	0.22	0.79
3, 9	1	0.28	0.11	0.28	0.11	6, 12	1	0.24	0.23	0.22	0.06
3, 9	2	0.33	0.05	0.33	0.05	6, 12	2	0.20	0.17	0.12	0.04
3, 9	3	0.23	0.05	0.23	0.05	6, 12	3	0.34	0.38	0.44	0.11

Table 6. Data illustration polytomous data: model comparison for exploratory LCMs.

C	#np	AIC	BIC
2	72	25,082	25,440
3	107	24,616	25,151
4	143	**24,431**	**25,148**
5	179	24,439	25,337
6	215	24,444	25,524

Note: C = number of classes; #np = number of estimated parameters.

Table 7. Data illustration polytomous data: model comparison for different RLCMs with four classes.

Appr.	Fused	Equation	C	λ_1	λ_2	#np	#nreg	BIC
R1	Class	(10)	4	0.31	—	84	63	24,982
R2	Cat	(11)	4	—	0.18	68	79	**24,689**
R3	Cat and Class	(12)	4	0.40	0.15	44	103	24,836
R4	Grouped Cat	(13)	4	0.45	—	82	65	24,777
R5	Grouped Class	(14)	4	0.65	—	79	67	24,776

Note: Appr. = approach; Cat = category; Eq. = equation for regularization penalty in Section 3.2; C = number of classes; #np = number of estimated parameters; #nreg = number of regularized item parameters.

For the best-fitting model of Approach R2 (i.e., fusing among categories), estimated class probabilities were 0.29, 0.22, 0.10, and 0.39, respectively. Estimated item response probabilities from this model are shown in Table 8. It can be seen that model estimation was quite successful in identifying parameters that were equal in the data generating model.

Table 8. Data illustration polytomous data: estimated item response probabilities in the RLCM with $C = 4$ classes and fused regularization among classes based on the minimal BIC.

Item	Cat	Class			
		1	2	3	4
1	0	0.07	0.79	0.82	0.82
1	1	0.31	0.07	0.06	0.06
1	2	0.31	0.07	0.06	0.06
1	3	0.31	0.07	0.06	0.06
2	0	0.22	0.85	0.87	0.91
2	1	0.26	0.05	0.01	0.06
2	2	0.26	0.05	0.01	0.03
2	3	0.26	0.05	0.11	0.00
3	0	0.16	0.74	0.19	0.80
3	1	0.28	0.16	0.27	0.10
3	2	0.28	0.05	0.27	0.05
3	3	0.28	0.05	0.27	0.05
4	0	0.26	0.88	0.28	0.85
4	1	0.36	0.05	0.24	0.04
4	2	0.19	0.02	0.24	0.02
4	3	0.19	0.05	0.24	0.09
5	0	0.10	0.13	0.46	0.92
5	1	0.30	0.29	0.18	0.02
5	2	0.30	0.29	0.18	0.02
5	3	0.30	0.29	0.18	0.04
6	0	0.22	0.24	0.30	0.76
6	1	0.26	0.18	0.32	0.07
6	2	0.26	0.18	0.06	0.03
6	3	0.26	0.40	0.32	0.14
7	0	0.07	0.79	0.85	0.85
7	1	0.31	0.09	0.05	0.05
7	2	0.31	0.09	0.05	0.05
7	3	0.31	0.03	0.05	0.05
8	0	0.25	0.86	0.91	0.88
8	1	0.25	0.06	0.03	0.06
8	2	0.25	0.06	0.03	0.06
8	3	0.25	0.02	0.03	0.00
9	0	0.13	0.82	0.10	0.82
9	1	0.29	0.06	0.30	0.06
9	2	0.29	0.06	0.30	0.06
9	3	0.29	0.06	0.30	0.06
10	0	0.20	0.89	0.37	0.82
10	1	0.42	0.05	0.25	0.05
10	2	0.19	0.01	0.13	0.03
10	3	0.19	0.05	0.25	0.10
11	0	0.10	0.16	0.55	0.91
11	1	0.30	0.28	0.15	0.03
11	2	0.30	0.28	0.15	0.03
11	3	0.30	0.28	0.15	0.03
12	0	0.24	0.20	0.20	0.76
12	1	0.21	0.21	0.35	0.10
12	2	0.21	0.21	0.10	0.04
12	3	0.34	0.38	0.35	0.10

Note: Cat = category.

5. Application of the SPM-LS Data

In this section, we illustrate the use of RLCM to the SPM-LS dataset.

5.1. Method

According to the topic of this special issue, the publicly available dataset from the Myszkowski and Storme (2018) study was reanalyzed. The original study compared various parametric item response models (i.e., 1PL, 2PL, 3PL, 4PL, and nested logit model) performed on a dataset comprised of $N = 499$ students (214 males and 285 females) aged between 19 and 24. The analyzed data consisted of responses on the 12 most difficult SPM items and are made freely available at https://data.mendeley.com/datasets/h3yhs5gy3w/1. For details regarding the data gathering procedure, we refer to Myszkowski and Storme (2018).

Each of the $I = 12$ items had one correct category and $K_i = 7$ distractors. To be consistent with the notation introduced in the paper and to ease interpretation of the results, we recoded the original dataset when using polytomous item responses. First, we scored the correct response as Category 0. Second, we recoded the order of distractors according to their frequency. In more detail, Category 1 in our rescored dataset was the most attractive distractor (i.e., most frequent distractor), while Category 7 was the least attractive distractor. The relative frequencies and references to categories of the original dataset are shown in Table 9. It could be supposed that there some especially attractive distractors for each item. However, many item category frequencies turned out to be relatively similar such that it could be that they would also function homogeneously among latent classes. We also analyzed the SPM-LS dataset in its dichotomous version in which Category 1 was scored as correct, and Category 0 summarized responses of all distractors.

Table 9. The last series of Raven's standard progressive matrices (SPM-LS) polytomous data: percentage frequencies and recoding table.

Item	Cat0	Cat1	Cat2	Cat3	Cat4	Cat5	Cat6	Cat7
SPM1	76.0 (7)	13.6 (3)	3.0 (1)	2.4 (4)	2.2 (6)	2.0 (2)	0.8 (5)	—
SPM2	91.0 (6)	3.0 (3)	2.4 (4)	2.2 (1)	0.8 (5)	0.4 (7)	0.2 (2)	—
SPM3	80.4 (8)	8.0 (2)	4.2 (6)	2.0 (4)	1.8 (3)	1.6 (5)	1.2 (7)	0.8 (1)
SPM4	82.4 (2)	5.6 (3)	3.2 (5)	2.6 (1)	2.2 (8)	1.8 (6)	1.2 (7)	1.0 (4)
SPM5	85.6 (1)	3.8 (2)	3.0 (3)	2.6 (7)	1.8 (6)	1.6 (5)	1.0 (4)	0.6 (8)
SPM6	76.4 (5)	7.0 (4)	5.2 (6)	3.0 (3)	2.8 (7)	2.6 (8)	2.0 (2)	1.0 (1)
SPM7	70.1 (1)	6.6 (4)	5.8 (5)	5.4 (3)	4.4 (8)	3.4 (6)	2.4 (7)	1.8 (2)
SPM8	58.1 (6)	7.6 (1)	7.0 (3)	6.6 (8)	6.4 (2)	6.2 (5)	5.8 (7)	2.2 (4)
SPM9	57.3 (3)	12.0 (5)	9.0 (1)	7.2 (4)	6.6 (8)	4.0 (7)	3.0 (2)	0.8 (6)
SPM10	39.5 (2)	17.2 (6)	11.2 (7)	8.0 (3)	7.8 (8)	7.4 (5)	6.0 (4)	2.8 (1)
SPM11	35.7 (4)	14.0 (1)	13.8 (7)	9.8 (5)	9.4 (6)	8.0 (3)	6.6 (2)	2.6 (8)
SPM12	32.5 (5)	15.4 (2)	14.2 (3)	10.4 (1)	8.2 (4)	8.2 (7)	7.4 (6)	3.6 (8)

Note: Numbers in parentheses denote the original item category.

For the dataset with dichotomous items, the exploratory LCM and the RLCM were fitted for two to six latent classes. Model selection was conducted based on the BIC. For the dataset with rescored polytomous items, we used the same number of classes for estimating the exploratory LCM. For the RLCM, we applied the fused regularization approach with respect to classes (Section 3.2.1), categories (Section 3.2.2), and to classes and categories in a simultaneous manner (Section 3.2.3).

5.2. Results

5.2.1. Results for Dichotomous Item Responses

For the SPM-LS dataset with dichotomous items, a series of exploratory LCMs and RLCMs with two to six classes was fitted. According to the BIC presented in Table 10, an exploratory LCM with four classes would be selected. When RLCMs were fitted, a model with five classes would be selected that had 19 regularized item parameters.

In Table 11, item response probabilities and skill class probabilities for the RLCM with $C = 5$ classes are shown. By considering the average item response probabilities per skill class $\overline{p}_{\bullet c} = (\sum_{i=1}^{I} p_{ic})/I$, Class C1 (12% frequency) was the least performing and Class C5 (37% frequency) the best performing class. Class C3 (40% frequency) could be seen as an intermediate class. Classes C2 and C4 were relatively rare. Compared to the medium Class C3, students in Class C2 had a particularly bad performance at Items 3, 6, and 11, but outperformed them on Items 7, 8, and 12. Students in Class C4 showed perfect performance on Items 8 and 9, but notably worse performance on Items 10 and 11. Interestingly, one could define a partial order on the classes if we allowed at most two violations of inequality conditions. In Figure 4, this partial order is depicted. The arrow from Class C1 to Class C3 means that C1 was smaller than C3. There are arrows with particular labels that indicate violations of the partial order. For example, C1 was approximately smaller than C2, and Items 3 and 11 violated the ordering property. To summarize, three out of the five classes fulfilled the ordering property for all items. Two classes possessed violations for two or three items and could be interpreted to detect subpopulations of subjects that showed latent differential item functioning.

Table 10. SPM-LS dichotomous data: model comparison for exploratory LCMs and RLCM.

	C	λ	#np	#nreg	BIC
LCM	2	0	25	0	5973
	3	0	38	0	5721
	4	0	51	0	**5680**
	5	0	64	0	5696
	6	0	77	0	5694
RLCM	2	0.01	25	0	5973
	3	0.33	35	3	5715
	4	0.38	39	12	5643
	5	0.29	45	19	**5621**
	6	0.53	45	32	5620

Note: C = number of classes; λ = regularization parameter of selected model with minimal BIC; #np = number of estimated parameters; #nreg = number of regularized parameters.

Table 11. SPM-LS dichotomous data: estimated item probabilities and latent class probabilities for best fitting RLCM with $C = 5$ latent classes.

Item	Class				
	C1	C2	C3	C4	C5
p_c	**0.12**	**0.04**	**0.40**	**0.07**	**0.37**
SPM1	0.39	0.39	0.83	0.83	0.83
SPM2	0.57	0.57	0.99	0.86	0.99
SPM3	0.33	0.00	0.86	0.96	0.96
SPM4	0.05	1.00	0.91	0.60	1.00
SPM5	0.08	1.00	0.96	0.77	1.00
SPM6	0.07	0.07	0.85	0.85	0.97
SPM7	0.20	0.83	0.58	0.83	0.95
SPM8	0.06	0.69	0.36	1.00	0.90
SPM9	0.16	0.34	0.34	1.00	0.90
SPM10	0.00	0.23	0.23	0.00	0.79
SPM11	0.14	0.00	0.14	0.00	0.77
SPM12	0.11	0.62	0.11	0.11	0.62
$\overline{p}_{\bullet c}$	0.18	0.48	0.60	0.65	0.89

Note: p_c = skill class probability; $\overline{p}_{\bullet c}$ = average of item probabilities within class c.

5.2.2. Results for Polytomous Item Responses

We now only briefly discuss the findings for the analysis of the SPM-LS dataset based on polytomous item responses. For the exploratory latent class models, the model with just two latent classes would be selected according to the BIC. However, the model with six latent classes would be selected according to the AIC. Given a large number of estimated item parameters, applying the RLCM seems to be required for obtaining a parsimonious model. The best-fitting model was obtained with $C = 3$ classes by fusing categories with a regularization parameter of $\lambda_2 = 0.24$. Classes C1 (28% frequency) and C2 (5% frequency) had low performance, while Class C3 was the high-performing class (67% frequency). As an illustration, we provide in Table 12 estimated item probabilities for the last three items. It can be seen that some of the categories were fused such that they had equal item response probabilities within a latent class. All item parameters are shown in Table A2 in Appendix B.

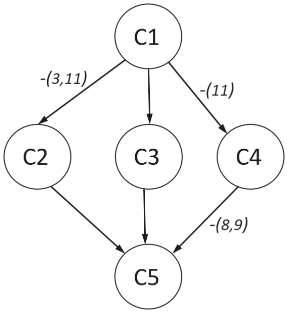


Figure 4. SPM-LS dichotomous data: partial order for latent class from RLCM.

Table 12. SPM-LS polytomous data: estimated item response probabilities and latent class probabilities for best-fitting RLCM with $C = 3$ latent classes for items SPM10, SPM11 and SPM12.

Item	Cat	C1	C2	C3
SPM10	0	0.08	0.00	0.56
SPM10	1	0.14	0.12	0.19
SPM10	2	0.26	0.40	0.03
SPM10	3	0.10	0.08	0.07
SPM10	4	0.13	0.00	0.06
SPM10	5	0.10	0.08	0.06
SPM10	6	0.10	0.32	0.03
SPM10	7	0.09	0.00	0.00

Item	Cat	C1	C2	C3
SPM11	0	0.11	0.26	0.48
SPM11	1	0.18	0.43	0.10
SPM11	2	0.21	0.00	0.12
SPM11	3	0.15	0.12	0.07
SPM11	4	0.14	0.00	0.08
SPM11	5	0.08	0.19	0.07
SPM11	6	0.07	0.00	0.07
SPM11	7	0.06	0.00	0.01

Item	Cat	C1	C2	C3
SPM12	0	0.03	0.24	0.45
SPM12	1	0.14	0.09	0.17
SPM12	2	0.19	0.44	0.10
SPM12	3	0.23	0.03	0.06
SPM12	4	0.12	0.00	0.07
SPM12	5	0.14	0.00	0.06
SPM12	6	0.07	0.20	0.07
SPM12	7	0.08	0.00	0.02

Note: Cat = category.

At the time of writing, results for polytomous data for the SPM-LS do not seem to be very consistent with those for dichotomous data. It could be the large number of parameters to be estimated (several hundred depending on the number of classes) for the relatively small sample size of $N = 499$ is critical. Other research has also shown that regularization methods for LCMs need sample sizes of at least 1000 or even more for performing satisfactorily (Chen et al. 2015).

6. Discussion

In this article, we proposed an extension of regularized latent class analysis to polytomous item responses. We have shown using the simulated data illustration and the SPM-LS dataset that fusing among classes or categories can be beneficial in terms of model parsimony and interpretation. Often, conceptualizing substantive questions as latent classes led researchers to easier to think in types of persons. This interpretation is not apparent in latent variables with continuous latent variables.

In our regularization approach to polytomous data, we based regularization penalties on distractors of items. Hence, the correct item response serves as a reference category. In LCA applications in which the definition of a reference category cannot be done, the regularization approach has certainly to be modified. Note that for $K_i + 1$ categories, only K_i item parameters per class can be independently estimated. Alternatively, a sum constraint $\sum_{k=0}^{K_i} \gamma_{ikc} = 0$ could be posed if γ_{ikc} ($k = 0, 1, \ldots, K_i$ denotes the item parameters of item i of category k in class c. Such constraints can be replaced by adding ridge-type penalties of the form $\lambda_3 \sum_{k=0}^{K_i} \gamma_{ikc}^2$ to the fused regularization penalty, where λ_3 is another regularization parameter. By squaring item parameters in the penalty function, they are uniformly shrunk to zero in the estimation.

By treating the correct item response as the reference category, regularization only operates on the categories for the incorrect response. As pointed out by an anonymous reviewer, it could be more appropriate by fusing classes for the correct item response and for incorrect item response categories separately. This would lead to an overidentified model because all class-specific item response probabilities would appear in the model. However, if, again, a ridge-type would be employed, the identification issue would disappear.

As the application of the regularization technique to an LCM results in a particular restricted LCM, it has to be shown that the model parameters can be identified. The analysis of necessary and sufficient conditions for identification in restricted LCMs was currently investigated (Gu and Xu 2018; Xu 2017). Because the inclusion of the penalty function, accompanied by a regularization parameter, introduces an additional amount of information in the estimation, it is unclear whether identifiability should be studied only on the likelihood part of the optimization function (see San Martín 2018 for a related discussion in Bayesian estimation).

It should be noted that similar regularization approaches have been studied for cognitive diagnostic models (Chen et al. 2015; Gu and Xu 2019; Liu and Kang 2019; Xu and Shang 2018). These kinds of models pose measurement models on D dichotomous latent variables. These D latent variables constitute 2^D latent classes. In addition, in this model class, the modeling of violations of the local independence assumption in LCA has been of interest (Kang et al. 2017; Tamhane et al. 2010).

Previous articles on the SPM-LS dataset also used distractor information by employing the nested logit model (NLM; Myszkowski and Storme 2018). The NLM is also very data-hungry, given the low sample size of the dataset. It has been argued that reliability can be increased by using distractor information (Myszkowski and Storme 2018; Storme et al. 2019). It should be noted that this is only true to the extent that item parameters can be reliably estimated. For $N = 499$ in the SPM-LS dataset, this will probably be not the case. Regularized estimation approaches could circumvent estimation issues (see Battauz 2019 for a similar approach in the nominal response model).

Finally, we would like to emphasize that the regularization approaches can be interpreted as empirical Bayesian approaches that employ hierarchical prior distributions on item parameters (van Erp et al. 2019). It can be expected that Bayesian variants of RLCMs are competitive to EM-based estimation, especially for small(er) samples.

Funding: This research received no external funding.

Acknowledgments: I am incredibly grateful to Nils Myszkowski for the long-standing patience and the invitation for contributing to the special issue 'Analysis of an Intelligence Dataset' in the *Journal of Intelligence*. We thank the three anonymous reviewers whose comments helped improve and clarify this manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AIC	Akaike information criterion
BIC	Bayesian information criterion
EM	expectation maximization
LASSO	least absolute shrinkage and selection operator
LCA	latent class analysis
LCM	latent class model
NLM	nested logit model
RLCA	regularized latent class analysis
RLCM	restricted latent class model
SPM-LS	last series of Raven's standard progressive matrices

Appendix A. Additional Results for Simulated Data Illustration with Polytomous Item Responses

In Table A1, estimated item response probabilities for the exploratory LCM with four latent classes are shown.

Table A1. Data illustration polytomous data: estimated item response probabilities in the exploratory LCM with $C = 4$ classes.

Item	Cat	Class			
		1	2	3	4
1	0	0.08	0.78	0.80	0.82
1	1	0.33	0.08	0.07	0.07
1	2	0.31	0.05	0.05	0.06
1	3	0.29	0.08	0.08	0.06
2	0	0.20	0.84	0.89	0.91
2	1	0.30	0.06	0.00	0.06
2	2	0.28	0.06	0.01	0.03
2	3	0.23	0.03	0.10	0.00
3	0	0.15	0.76	0.20	0.80
3	1	0.25	0.15	0.34	0.10
3	2	0.36	0.06	0.18	0.05
3	3	0.24	0.03	0.28	0.06
4	0	0.27	0.89	0.30	0.86
4	1	0.35	0.04	0.28	0.04
4	2	0.19	0.02	0.21	0.01
4	3	0.19	0.05	0.22	0.09
5	0	0.10	0.11	0.45	0.92
5	1	0.34	0.27	0.23	0.02
5	2	0.32	0.31	0.17	0.02
5	3	0.24	0.31	0.15	0.04
6	0	0.22	0.24	0.31	0.77
6	1	0.23	0.17	0.16	0.07
6	2	0.23	0.21	0.06	0.03
6	3	0.31	0.39	0.47	0.13
7	0	0.07	0.79	0.83	0.84
7	1	0.34	0.10	0.06	0.05
7	2	0.31	0.07	0.06	0.06
7	3	0.28	0.03	0.06	0.05
8	0	0.24	0.86	0.91	0.88
8	1	0.23	0.06	0.05	0.06
8	2	0.24	0.06	0.02	0.06
8	3	0.28	0.02	0.03	0.00
9	0	0.14	0.82	0.14	0.81
9	1	0.27	0.08	0.26	0.08
9	2	0.32	0.05	0.33	0.05
9	3	0.27	0.05	0.27	0.05
10	0	0.19	0.89	0.40	0.83
10	1	0.42	0.05	0.28	0.05
10	2	0.19	0.01	0.12	0.03
10	3	0.20	0.05	0.20	0.10
11	0	0.10	0.13	0.55	0.90
11	1	0.32	0.28	0.10	0.03
11	2	0.26	0.34	0.17	0.03
11	3	0.32	0.25	0.18	0.04
12	0	0.25	0.19	0.21	0.77
12	1	0.23	0.24	0.27	0.08
12	2	0.19	0.20	0.10	0.04
12	3	0.34	0.38	0.42	0.12

Note: Cat = category.

Appendix B. Additional Results for SPM-LS Dataset with Polytomous Item Responses

Table A2 shows all estimated item response probabilities for the SPM-LS dataset with polytomous items for the best fitting RLCM with $C = 3$ classes.

Table A2. SPM-LS polytomous data: estimated item response probabilities and latent class probabilities for best-fitting RLCM with $C = 3$ latent classes.

Item	Cat	C1	C2	C3
SPM1	0	0.73	0.15	0.82
SPM1	1	0.11	0.48	0.12
SPM1	2	0.06	0.01	0.02
SPM1	3	0.07	0.04	0.01
SPM1	4	0.00	0.24	0.01
SPM1	5	0.01	0.08	0.02
SPM1	6	0.02	0.00	0.00
SPM1	7	0.00	0.00	0.00
SPM2	0	0.87	0.32	0.97
SPM2	1	0.02	0.12	0.03
SPM2	2	0.02	0.36	0.00
SPM2	3	0.07	0.00	0.00
SPM2	4	0.01	0.12	0.00
SPM2	5	0.00	0.08	0.00
SPM2	6	0.01	0.00	0.00
SPM2	7	0.00	0.00	0.00
SPM3	0	0.67	0.04	0.92
SPM3	1	0.17	0.00	0.05
SPM3	2	0.08	0.40	0.00
SPM3	3	0.01	0.32	0.00
SPM3	4	0.01	0.04	0.02
SPM3	5	0.00	0.12	0.01
SPM3	6	0.04	0.04	0.00
SPM3	7	0.02	0.04	0.00
SPM4	0	0.62	0.00	0.98
SPM4	1	0.11	0.40	0.01
SPM4	2	0.04	0.36	0.00
SPM4	3	0.07	0.08	0.00
SPM4	4	0.05	0.00	0.01
SPM4	5	0.04	0.12	0.00
SPM4	6	0.04	0.00	0.00
SPM4	7	0.03	0.04	0.00
SPM5	0	0.72	0.04	0.98
SPM5	1	0.05	0.48	0.00
SPM5	2	0.03	0.32	0.01
SPM5	3	0.08	0.00	0.00
SPM5	4	0.06	0.00	0.00
SPM5	5	0.02	0.08	0.01
SPM5	6	0.02	0.08	0.00
SPM5	7	0.02	0.00	0.00
SPM6	0	0.51	0.08	0.92
SPM6	1	0.10	0.36	0.04
SPM6	2	0.11	0.00	0.03
SPM6	3	0.09	0.00	0.01
SPM6	4	0.06	0.24	0.00
SPM6	5	0.06	0.20	0.00
SPM6	6	0.05	0.08	0.00
SPM6	7	0.02	0.04	0.00
SPM7	0	0.38	0.20	0.88
SPM7	1	0.06	0.12	0.06
SPM7	2	0.13	0.28	0.01
SPM7	3	0.12	0.28	0.01
SPM7	4	0.11	0.00	0.02
SPM7	5	0.07	0.00	0.02
SPM7	6	0.09	0.00	0.00
SPM7	7	0.04	0.12	0.00
SPM8	0	0.18	0.00	0.81
SPM8	1	0.24	0.04	0.01
SPM8	2	0.08	0.00	0.07
SPM8	3	0.14	0.08	0.03
SPM8	4	0.11	0.20	0.03
SPM8	5	0.07	0.28	0.04
SPM8	6	0.11	0.40	0.01
SPM8	7	0.07	0.00	0.00
SPM9	0	0.25	0.22	0.73
SPM9	1	0.14	0.03	0.12
SPM9	2	0.17	0.00	0.07
SPM9	3	0.12	0.32	0.03
SPM9	4	0.19	0.15	0.01
SPM9	5	0.06	0.00	0.03
SPM9	6	0.06	0.16	0.01
SPM9	7	0.01	0.12	0.00
SPM10	0	0.08	0.00	0.56
SPM10	1	0.14	0.12	0.19
SPM10	2	0.26	0.40	0.03
SPM10	3	0.10	0.08	0.07
SPM10	4	0.13	0.00	0.06
SPM10	5	0.10	0.08	0.06
SPM10	6	0.10	0.32	0.03
SPM10	7	0.09	0.00	0.00
SPM11	0	0.11	0.26	0.48
SPM11	1	0.18	0.43	0.10
SPM11	2	0.21	0.00	0.12
SPM11	3	0.15	0.12	0.07
SPM11	4	0.14	0.00	0.08
SPM11	5	0.08	0.19	0.07
SPM11	6	0.07	0.00	0.07
SPM11	7	0.06	0.00	0.01
SPM12	0	0.03	0.24	0.45
SPM12	1	0.14	0.09	0.17
SPM12	2	0.19	0.44	0.10
SPM12	3	0.23	0.03	0.06
SPM12	4	0.12	0.00	0.07
SPM12	5	0.14	0.00	0.06
SPM12	6	0.07	0.20	0.07
SPM12	7	0.08	0.00	0.02

Note: Cat = category.

References

Agresti, Alan, and Maria Kateri. 2014. Some remarks on latent variable models in categorical data analysis. *Communications in Statistics Theory and Methods* 43: 801–14. [CrossRef]

Battauz, Michela. 2019. Regularized estimation of the nominal response model. *Multivariate Behavioral Research*. [CrossRef] [PubMed]

Bhattacharya, Sakyajit, and Paul D. McNicholas. 2014. A LASSO-penalized BIC for mixture model selection. *Advances in Data Analysis and Classification* 8: 45–61. [CrossRef]

Borsboom, Denny, Mijke Rhemtulla, Angelique O. J. Cramer, Han L. J. van der Maas, Marten Scheffer, and Conor V. Dolan. 2016. Kinds versus continua: A review of psychometric approaches to uncover the structure of psychiatric constructs. *Psychological Medicine* 46: 1567–79. [CrossRef]

Cao, Peng, Xiaoli Liu, Hezi Liu, Jinzhu Yang, Dazhe Zhao, Min Huang, and Osmar Zaiane. 2018. Generalized fused group lasso regularized multi-task feature learning for predicting cognitive outcomes in Alzheimers disease. *Computer Methods and Programs in Biomedicine* 162: 19–45. [CrossRef]

Chen, Yunxiao, Jingchen Liu, Gongjun Xu, and Zhiliang Ying. 2015. Statistical analysis of Q-matrix based diagnostic classification models. *Journal of the American Statistical Association* 110: 850–66. [CrossRef]

Chen, Yunxiao, Xiaoou Li, Jingchen Liu, and Zhiliang Ying. 2017. Regularized latent class analysis with application in cognitive diagnosis. *Psychometrika* 82: 660–92. [CrossRef]

Chen, Yunxiao, Xiaoou Li, Jingchen Liu, and Zhiliang Ying. 2018. Robust measurement via a fused latent and graphical item response theory model. *Psychometrika* 83: 538–62. [CrossRef]

Collins, Linda M., and Stephanie T. Lanza. 2009. *Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral, and Health Sciences*. New York: Wiley. [CrossRef]

DeSantis, Stacia M., E. Andrés Houseman, Brent A. Coull, Catherine L. Nutt, and Rebecca A. Betensky. 2012. Supervised Bayesian latent class models for high-dimensional data. *Statistics in Medicine* 31: 1342–60. [CrossRef]

DeSantis, Stacia M., E. Andrés Houseman, Brent A. Coull, Anat Stemmer-Rachamimov, and Rebecca A. Betensky. 2008. A penalized latent class model for ordinal data. *Biostatistics* 9: 249–62. [CrossRef]

Fan, Jianqing, and Runze Li. 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 96: 1348–60. [CrossRef]

Finch, W. Holmes, and Kendall C. Bronk. 2011. Conducting confirmatory latent class analysis using Mplus. *Structural Equation Modeling* 18: 132–51. [CrossRef]

Fop, Michael, and Thomas B. Murphy. 2018. Variable selection methods for model-based clustering. *Statistics Surveys* 12: 18–65. [CrossRef]

Formann, Anton K. 1982. Linear logistic latent class analysis. *Biometrical Journal* 24: 171–90. [CrossRef]

Formann, Anton K. 1992. Linear logistic latent class analysis for polytomous data. *Journal of the American Statistical Association* 87: 476–86. [CrossRef]

Formann, Anton K. 2007. (Almost) equivalence between conditional and mixture maximum likelihood estimates for some models of the Rasch type. In *Multivariate and Mixture Distribution Rasch Models*. Edited by Matthias von Davier and Claus H. Carstensen. New York: Springer, pp. 177–89. [CrossRef]

Formann, Anton K., and Thomas Kohlmann. 1998. Structural latent class models. *Sociological Methods & Research* 26: 530–65. [CrossRef]

George, Ann C., Alexander Robitzsch, Thomas Kiefer, Jürgen Groß, and Ali Ünlü. 2016. The R package CDM for cognitive diagnosis models. *Journal of Statistical Software* 74: 1–24. [CrossRef]

Gu, Yuqi, and Gongjun Xu. 2018. Partial identifiability of restricted latent class models. *arXiv* arXiv:1803.04353. [CrossRef]

Gu, Yuqi, and Gongjun Xu. 2019. Learning attribute patterns in high-dimensional structured latent attribute models. *Journal of Machine Learning Research* 20: 115.

Hastie, Trevor, Robert Tibshirani, and Martin Wainwright. 2015. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Boca Raton: CRC Press. [CrossRef]

Houseman, E. Andrés, Brent A. Coull, and Rebecca A. Betensky. 2006. Feature-specific penalized latent class analysis for genomic data. *Biometrics* 62: 1062–70. [CrossRef]

Huang, Jian, Patrick Breheny, and Shuangge Ma. 2012. A selective review of group selection in high-dimensional models. *ss* 27: 481–99. [CrossRef] [PubMed]

Huang, Po-Hsien, Hung Chen, and Li-Jen Weng. 2017. A penalized likelihood method for structural equation modeling. *Psychometrika* 82: 329–54. [CrossRef] [PubMed]

Jacobucci, Ross, Kevin J. Grimm, and John J. McArdle. 2016. Regularized structural equation modeling. *Structural Equation Modeling* 23: 555–66. [CrossRef] [PubMed]

Janssen, Anne B., and Christian Geiser. 2010. On the relationship between solution strategies in two mental rotation tasks. *Learning and Individual Differences* 20: 473–78. [CrossRef]

Kang, Hyeon-Ah, Jingchen Liu, and Zhiliang Ying. 2017. A graphical diagnostic classification model. *arXiv* arXiv:1707.06318.

Keribin, Christine. 2000. Consistent estimation of the order of mixture models. *Sankhyā: The Indian Journal of Statistics, Series A* 62: 49–66.

Langeheine, Rolf, and Jürgen Rost. 1988. *Latent Trait and Latent Class Models*. New York: Plenum Press. [CrossRef]

Lazarsfeld, Paul F., and Neil W. Henry. 1968. *Latent Structure Analysis*. Boston: Houghton Mifflin.

Leoutsakos, Jeannie-Marie S., Karen Bandeen-Roche, Elzabeth Garrett-Mayer, and Peter P. Zandi. 2011. Incorporating scientific knowledge into phenotype development: Penalized latent class regression. *Statistics in Medicine* 30: 784–98. [CrossRef]

Liu, Jingchen, and Hyeon-Ah Kang. 2019. Q-matrix learning via latent variable selection and identifiability. In *Handbook of Diagnostic Classification Models*. Edited by Matthias von Davier and Young-Sun Lee. Cham: Springer, pp. 247–63. [CrossRef]

Liu, Xiaoli, Peng Cao, Jianzhong Wang, Jun Kong, and Dazhe Zhao. 2019. Fused group lasso regularized multi-task feature learning and its application to the cognitive performance prediction of Alzheimer's disease. *Neuroinformatics* 17: 271–94. [CrossRef]

Myszkowski, Nils, and Martin Storme. 2018. A snapshot of g. Binary and polytomous item-response theory investigations of the last series of the standard progressive matrices (SPM-LS). *Intelligence* 68: 109–16. [CrossRef]

Nussbeck, Fritjof W., and Michael Eid. 2015. Multimethod latent class analysis. *Frontiers in Psychology* 6: 1332. [CrossRef]

Oberski, Daniel L., Jacques A. P. Hagenaars, and Willem E. Saris. 2015. The latent class multitrait-multimethod model. *Psychological Methods* 20: 422–43. [CrossRef] [PubMed]

Oelker, Margret-Ruth, and Gerhard Tutz. 2017. A uniform framework for the combination of penalties in generalized structured models. *Advances in Data Analysis and Classification* 11: 97–120. [CrossRef]

Robitzsch, Alexander. 2020. sirt: Supplementary Item Response Theory Models. R Package Version 3.9-4. Available online: https://CRAN.R-project.org/package=sirt (accessed on 17 February 2020).

Robitzsch, Alexander, and Ann C. George. 2019. The R package CDM for diagnostic modeling. In *Handbook of Diagnostic Classification Models*. Edited by Matthias von Davier and Young-Sun Lee. Cham: Springer, pp. 549–72. [CrossRef]

Ruan, Lingyan, Ming Yuan, and Hui Zou. 2011. Regularized parameter estimation in high-dimensional Gaussian mixture models. *Neural Computation* 23: 1605–22. [CrossRef] [PubMed]

San Martín, Ernesto. 2018. Identifiability of structural characteristics: How relevant is it for the Bayesian approach? *Brazilian Journal of Probability and Statistics* 32: 346–73. [CrossRef]

Scharf, Florian, and Steffen Nestler. 2019. Should regularization replace simple structure rotation in exploratory factor analysis? *Structural Equation Modeling* 26: 576–90. [CrossRef]

Schmiege, Sarah J., Katherine E. Masyn, and Angela D. Bryan. 2018. Confirmatory latent class analysis: Illustrations of empirically driven and theoretically driven model constraints. *Organizational Research Methods* 21: 983–1001. [CrossRef]

Storme, Martin, Nils Myszkowski, Simon Baron, and David Bernard. 2019. Same test, better scores: Boosting the reliability of short online intelligence recruitment tests with nested logit item response theory models. *Journal of Intelligence* 7: 17. [CrossRef]

Sun, Jianan, Yunxiao Chen, Jingchen Liu, Zhiliang Ying, and Tao Xin. 2016. Latent variable selection for multidimensional item response theory models via L_1 regularization. *Psychometrika* 81: 921–39. [CrossRef]

Sun, Jiehuan, Jose D. Herazo-Maya, Philip L. Molyneaux, Toby M. Maher, Naftali Kaminski, and Hongyu Zhao. 2019. Regularized latent class model for joint analysis of high-dimensional longitudinal biomarkers and a time-to-event outcome. *Biometrics* 75: 69–77. [CrossRef]

Tamhane, Ajit C., Dingxi Qiu, and Bruce E. Ankenman. 2010. A parametric mixture model for clustering multivariate binary data. *Statistical Analysis and Data Mining* 3: 3–19. [CrossRef]

Tibshirani, Robert, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. 2005. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 67: 91–108. [CrossRef]

Tutz, Gerhard, and Jan Gertheiss. 2016. Regularized regression for categorical data. *Statistical Modelling* 16: 161–200. [CrossRef]

Tutz, Gerhard, and Gunther Schauberger. 2015. A penalty approach to differential item functioning in Rasch models. *Psychometrika* 80: 21–43. [CrossRef] [PubMed]

van Erp, Sara, Daniel L. Oberski, and Joris Mulder. 2019. Shrinkage priors for Bayesian penalized regression. *Journal of Mathematical Psychology* 89: 31–50. [CrossRef]

von Davier, Matthias. 2008. A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology* 61: 287–307. [CrossRef] [PubMed] von Davier, Matthias. 2010. Hierarchical mixtures of diagnostic models. *Psychological Test and Assessment Modeling* 52: 8–28.

von Davier, Matthias, and Young-Sun Lee, eds. 2019. *Handbook of Diagnostic Classification Models*. Cham: Springer. [CrossRef]

von Davier, Matthias, Bobby Naemi, and Richard D. Roberts. 2012. Factorial versus typological models: A comparison of methods for personality data. *Measurement: Interdisciplinary Research and Perspectives* 10: 185–208. [CrossRef]

Wang, Chun, and Jing Lu. 2020. Learning attribute hierarchies from data: Two exploratory approaches. *Journal of Educational and Behavioral Statistics*. [CrossRef]

Wu, Baolin. 2013. Sparse cluster analysis of large-scale discrete variables with application to single nucleotide polymorphism data. *Journal of Applied Statistics* 40: 358–67. [CrossRef]

Wu, Zhenke, Livia Casciola-Rosen, Antony Rosen, and Scott L. Zeger. 2018. A Bayesian approach to restricted latent class models for scientifically-structured clustering of multivariate binary outcomes. *arXiv* arXiv:1808.08326.

Xu, Gongjun. 2017. Identifiability of restricted latent class models with binary responses. *Annals of Statistics* 45: 675–707. [CrossRef]

Xu, Gongjun, and Zhuoran Shang. 2018. Identifying latent structures in restricted latent class models. *Journal of the American Statistical Association* 113: 1284–95. [CrossRef]

Yamamoto, Michio, and Kenichi Hayashi. 2015. Clustering of multivariate binary data with dimension reduction via L_1-regularized likelihood maximization. *Pattern Recognition* 48: 3959–68. [CrossRef]

Zhang, Cun-Hui. 2010. Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics* 38: 894–942. [CrossRef]

MDPI
St. Alban-Anlage 66
4052 Basel
Switzerland
Tel. +41 61 683 77 34
Fax +41 61 302 89 18
www.mdpi.com

Journal of Intelligence Editorial Office
E-mail: jintelligence@mdpi.com
www.mdpi.com/journal/jintelligence

MDPI
St. Alban-Anlage 66
4052 Basel
Switzerland

Tel: +41 61 683 77 34
Fax: +41 61 302 89 18

www.mdpi.com

ISBN 978-3-0365-0040-9