Springer Proceedings in Mathematics & Statistics

Marie Wiberg
Steven Culpepper
Rianne Janssen
Jorge González
Dylan Molenaar *Editors*

Quantitative Psychology

83rd Annual Meeting of the Psychometric Society, New York, NY 2018

Springer

Springer Proceedings in Mathematics & Statistics

Volume 265

Springer Proceedings in Mathematics & Statistics

This book series features volumes composed of selected contributions from workshops and conferences in all areas of current research in mathematics and statistics, including operation research and optimization. In addition to an overall evaluation of the interest, scientific quality, and timeliness of each proposal at the hands of the publisher, individual contributions are all refereed to the high quality standards of leading journals in the field. Thus, this series provides the research community with well-edited, authoritative reports on developments in the most exciting areas of mathematical and statistical research today.

More information about this series at http://www.springer.com/series/10533

Marie Wiberg · Steven Culpepper ·
Rianne Janssen · Jorge González ·
Dylan Molenaar
Editors

Quantitative Psychology

83rd Annual Meeting of the Psychometric Society, New York, NY 2018

Springer

Editors
Marie Wiberg
Department of Statistics, Umeå School of Business, Economics and Statistics
Umeå University
Umeå, Sweden

Rianne Janssen
Faculty of Psychology and Educational Sciences
KU Leuven
Leuven, Belgium

Dylan Molenaar
Department of Psychology
University of Amsterdam
Amsterdam, The Netherlands

Steven Culpepper
Department of Statistics
University of Illinois at Urbana-Champaign
Champaign, IL, USA

Jorge González
Facultad de Matematicas
Pontificia Universidad Catolica de Chile
Santiago, Chile

ISSN 2194-1009 ISSN 2194-1017 (electronic)
Springer Proceedings in Mathematics & Statistics
ISBN 978-3-030-01309-7 ISBN 978-3-030-01310-3 (eBook)
https://doi.org/10.1007/978-3-030-01310-3

Mathematics Subject Classification (2010): 62P15, 62-06, 62H12, 62-07

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

This volume represents presentations given at the 83rd annual meeting of the Psychometric Society, organized by Columbia University and held in New York, USA, during July 9–13, 2018. The meeting attracted 505 participants, and 286 papers were presented, of which 81 were part of a symposium. There were 106 poster presentations, 3 pre-conference workshops, 4 keynote presentations, 3 invited presentations, 2 career award presentations, 3 state-of-the-art presentations, 1 dissertation award winner, and 18 symposia.

Since the 77th meeting in Lincoln, Nebraska, Springer publishes the proceedings volume from the annual meeting of the Psychometric Society to allow presenters to make their ideas available quickly to the wider research community, while still undergoing a thorough review process. The first six volumes of the meetings in Lincoln, Arnhem, Madison, Beijing, Asheville, and Zurich were received successfully, and we expect a successful reception of these proceedings too.

We asked the authors to use their presentation at the meeting as the basis of their chapters, possibly extended with new ideas or additional information. The result is a selection of 38 state-of-the-art chapters addressing a diverse set of psychometric topics, including item response theory, multistage adaptive testing, and cognitive diagnostic models.

Umeå, Sweden	Marie Wiberg
Urbana-Champaign, IL, USA	Steven Culpepper
Leuven, Belgium	Rianne Janssen
Santiago, Chile	Jorge González
Amsterdam, The Netherlands	Dylan Molenaar

Contents

Explanatory Item Response Theory Models: Impact on Validity and Test Development?

Susan Embretson

Abstract Many explanatory item response theory (IRT) models have been developed since Fischer's (Acta Psychologica 37:359–374, 1973) linear logistic test model was published. However, despite their applicability to typical test data, actual impact on test development and validation has been limited. The purpose of this chapter is to explicate the importance of explanatory IRT models in the context of a framework that interrelates the five aspects of validity (Embretson in Educ Meas Issues Pract 35, 6–22, 2016). In this framework, the *response processes* aspect of validity impacts other aspects. Studies on a fluid intelligence test are presented to illustrate the relevancy of explanatory IRT models to validity, as well as to test development.

Keywords Item response theory · Explanatory models · Validity

1 Introduction

Since Fischer (1973) introduced the linear logistic test model (LLTM), many additional explanatory IRT models have been developed to estimate the impact of item complexity on item parameters. These models include the linear partial credit model (LPCM; Fischer & Ponocny, 1995), the linear logistic test model with response error term (LLTM-R; Janssen, Schepers, & Peres, 2004), the constrained two parameter logistic model (2PL-Constrained; Embretson, 1999) and the Rasch facet model (Linacre, 1989). Explanatory IRT models also can include covariates for both items and persons, as well as within-person interactions (De Boeck & Wilson, 2004). Several models can detect strategy differences between persons, such as mixture distribution models (Rost, 1990; Rost & von Davier, 1995) and mixed models that include response time to detect strategies (Molenaar & De Boeck, 2018). Further, hierarchical models can be used in an explanatory fashion, such as item family models (Glas, van der Linden & Geerlings, 2010) and a criterion-referenced model (Janssen, Tuerlinckx, Meulder & De Boeck, 2000). Multidimensional IRT models with defined

S. Embretson (✉)
Georgia Institute of Technology, Atlanta, GA 30328, USA
e-mail: susan.embretson@psych.gatech.edu

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 265, https://doi.org/10.1007/978-3-030-01310-3_1

dimensions, such as the bifactor MIRT (Reise, 2012) or the multicomponent latent trait model (MLTM; Embretson, 1984, 1997) also can be used as explanatory IRT models. The *Handbook of Item Response Theory* (van der Linden, 2016) includes several explanatory models. Janssen (2016) notes that explanatory IRT models have been applied to many tests, ranging from mathematics, reading and reasoning to personality and emotions.

However, despite the existence of these models for several decades and their applicability to typical test data, actual impact on test development and validation has been limited. The purpose of this chapter is to highlight the importance of explanatory IRT models in test development. Studies on the development of a fluid intelligence test are presented to illustrate the use of explanatory IRT models in test design and validation. Prior to presenting the studies, background on the validity concept and a framework that unifies the various aspects are presented.

1.1 Test Validity Framework

In the current *Standards for Educational and Psychological Testing* (2014), validity is conceptualized as a single type (construct validity) with five aspects. First, the *content* aspect of construct validity is the representation of skills, knowledge and attributes on the test. It is supported by specified test content, such as blueprints that define item skills, knowledge or attribute representation, as well as specifications of test administration and scoring conditions. Second, the *response processes* aspect of validity consists of evidence on the cognitive activities engaged in by the examinees. These cognitive activities are assumed to be essential to the meaning of the construct measured by a test. The *Standards for Educational and Psychological Testing* describes several direct methods to observe examinees' processing on test items, such as eye-trackers movements, videos and concurrent and retrospective verbal reports/observations, as well as response times to items or the whole test. Third, the *internal structure* aspect of construct validity includes psychometric properties of a test as relevant to the intended construct. Thus, internal consistency reliability, test dimensionality and differential item functioning (DIF) are appropriate types of evidence. Item selection, as part of test design, has a direct impact on internal structure. Fourth, the r*elationship to other variables* aspect concerns how the test relates to other traits and criteria, as well as to examinee background variables (i.e., demographics, prior experience, etc.). Evidence relevant to this aspect should be consistent with the goals of measurement. Fifth, the *consequences* aspect of validity concerns how test use has adverse impact on different groups of examinees. While the test may not have significant DIF, studies may nonetheless show that the test has adverse impact if used for selection or placement. Adverse impact is particularly detrimental to test quality if based on construct-irrelevant aspects of performance.

The various aspects of validity can be conceptualized as a unified system with causal interrelationships (Embretson, 2017). Figure 1 organizes the five aspects into two general areas, internal and external, which concern test

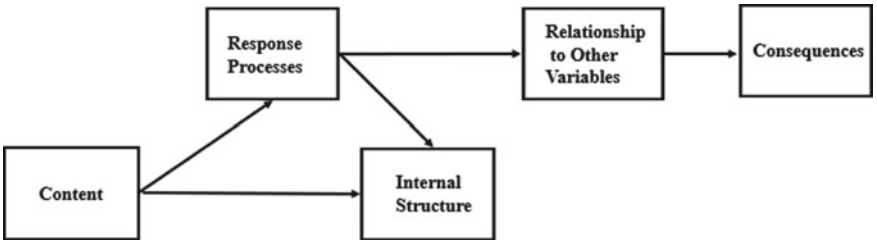

Fig. 1 Unified framework for validity

meaning and test significance, respectively. Thus, the *content*, *response processes* and *internal structure* aspects are relevant to defining the meaning of the construct while the *relationships to other variables* and *consequences* aspects define the significance of the test. Notice that the *content* and *response processes* aspect drive the other aspects causally in this framework. Importantly, these two aspects can be manipulated in test development. That is, item design, test specifications and testing conditions can impact test meaning. Thus, understanding the relationship between test content and response processes can be crucial in test development to measure the intended construct.

Unfortunately, the methods for understanding *response processes* described in the *Standards* have substantial limitations. Both eye-tracker data and talk aloud data are typically expensive to collect and analyze as well as impacting the nature of processing for examinees. Further, unless elaborated in the context of a model, the utility of response time data may be limited to identifying guessing or inappropriate responses. Importantly, explanatory IRT modeling can be applied to standard test data with no impact on examinees responses. Further, such models permit hypotheses to be tested about the nature of *response processes* through relationships of item content features and item responses.

2 Explanatory IRT Models in Item Design: Examples from ART

The Abstract Reasoning Test (ART) was developed in the context of research on *response processes*. ART is a test of fluid intelligence used to predict learning and performance in a variety of settings (Embretson, 2017). ART consists of matrix completion items as shown in Fig. 2. In these items, the examinee must identify the figure that completes the matrix based on the relationships between the figures across the rows and down the columns.

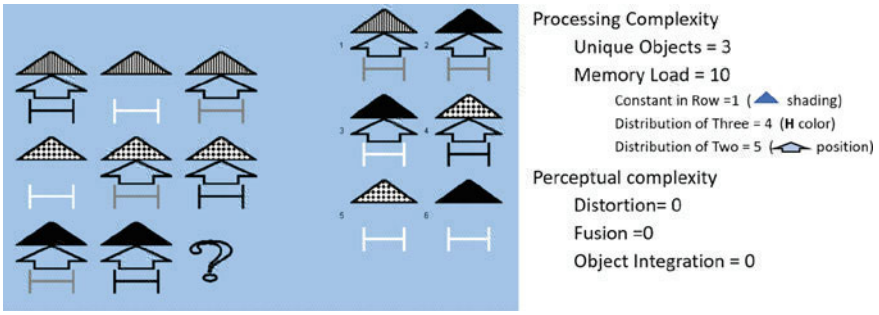

Fig. 2 Example of an ART item

2.1 Theory of Response Processes on Matrix Problems

Consistent with the Carpenter, Just and Shell's (1990) theory, it was hypothesized that examinees process the various elements individually in the matrix entries to find relationships. According to the theory, *processing complexity* is driven by the number of unique objects (as counted in the first entry) and memory load in finding relationships. Memory load depends on both the number and types of relationships, which are hypothesized to be ordered by complexity as follows: 1 = Constant in a Row (or column), the same figure appears in a row; 2 = Pairwise Progressions, figures change in the same way in each row; 3 = Figure Addition/Subtraction, the third column results from overlaying the first and second columns and subtracting common figures; 4 = Distribution of Three, a figure appears once and only once in each row and column and 5 = Distribution of Two, one figure is systematically missing in each row and column. Figure 2 illustrates relationships #1, #4 and #5 (see key on right) and Fig. 4 illustrates relationship #3. Relationship #2 could be illustrated by a change in object size across rows. Carpenter et al. (1990) postulate that these relationships are tried sequentially by examinees, such that Constant in a Row is considered before Pairwise Progressions and so forth. Thus, the Memory Load score is highest for the Distribution of Two relationships. Figure 2 shows numerical impact on Memory Load for three types of relationships. The difficulty of solving matrix problems also is hypothesized to depend on *perceptual complexity*, which is determined by Distortion, Fusion or Integration of objects in an entry. Figure 2 has none of these sources of *perceptual complexity* while Fig. 4 illustrates object integration in the matrix on the right side. Each matrix item can be scored for the *processing* and *perceptual complexity* variables. Item difficulty is postulated to result from these variables because they drive cognitive complexity.

2.2 Explanatory Modeling of Response Processes on ART Matrix Problems

An explanatory modeling of ART item difficulty results from applying LLTM to item response data, using the scores for matrix problem complexity. LLTM is given as follows:

$$P(\theta) = \frac{\exp(\theta_j - \sum_k \tau_k q_{ik} + \tau_0)}{1 + \exp(\theta_j - \sum_k \tau_k q_{ik} + \tau_0)} \quad (1)$$

where q_{ik} is the score for item *i* on attribute *k*, τ_k is the weight of attribute *k* in item difficulty and τ_0 is an intercept. Finally, θ_j is the ability of person *j*.

LLTM was applied to model item responses for ART items, scored for the two predictors of *processing complexity* and the three predictors of *perceptual complexity*. For example, a sample of 705 Air Force recruits were administered a form of ART with 30 items. The delta statistic, which is a likelihood ratio index of fit (Embretson, 1999) similar in magnitude to a multiple correlation, indicated that LLTM had strong fit to the data ($\Delta = .78$). The *processing complexity* variables had the strongest impact, especially memory load, which supports the theory.

2.3 Impact of Explanatory Modeling on Item Design for Matrix Problems

These results and the scoring system had direct impact on item and test design for ART. An automatic item generator was developed for ART items. Abstract structures were specified to define the objects within each cell of the 3 × 3 display and the response options. Types of relationships, as described above, specifies the changes in objects (e.g., circles, arrows, squares, etc.) and/or their properties (e.g., shading, borders, distortion, size, etc.) across columns and rows. LLTM results on military samples indicated high predictability of item difficulty by the generating structure ($\Delta = .90$) and continued prediction by the five variables defining cognitive complexity ($\Delta = .79$).

3 Strategy Modeling in Test Design: Example from ART

Examinee differences in item solving strategies and potential impact on the various aspects of validity was examined in two studies. In Study 1, ART was administered with the original brief instructions. In Study 2, ART was administered with an expanded version of the instructions with examples of each type of relationship. In both studies, strategies were examined through mixture modeling.

3.1 Mixture Modeling to Identify Latent Classes

The mixture Rasch model (Rost & von Davier, 1995) can be applied to identify classes of examinees that vary in item difficulty ordering, which is postulated to arise from applying different item solving strategies. The mixture Rasch model is given as follows:

$$P(\theta) = \Sigma_g \pi_g \frac{\exp(\theta_{jg} - \beta_{ig})}{1 + \exp(\theta_{jg} - \beta_{ig})} \quad (2)$$

where β_{ig} is the difficulty of item i in class g, θ_{jg} is the ability of person j in class g and π_g is the probability of class g. Classes are identified empirically to maximize model fit. However, class interpretation can be examined by follow-up explanatory modeling (e.g., applying LLTM within classes) or by comparing external correlates of ability.

3.2 Study 1

Method. A form of ART with 30 items was administered to 803 Air Force recruits who were completing basic training. The ART instructions concerned the nature of matrix problems as defined by relationships in the row and columns in the 3 × 3 matrices. However, the scope of relationships that could be involved was not covered. ART was administered without time limits. Item parameters were estimated with the Rasch model and with the mixture Rasch model. In both cases the mean item parameter was set to zero.

Results from other tests were available on the examinees, including the *Armed Services Vocational Aptitude Battery* (ASVAB).

Results. The test had moderate difficulty for the sample based on raw scores ($M = 18.097$, $SD = 5.784$) and latent trait estimates ($M = .636$, $SD = 1.228$). Racial-ethnic comparisons were between groups with $N > 50$. The latent trait estimates were significant ($F_{2,743} = 8.722$, $p < .001$, $\eta^2 = .023$). Standardized differences of ($d = .452$) for African Americans and ($d = .136$) for Hispanics were observed as compared to Caucasians.

The mixture Rasch model was applied with varying numbers of classes. Table 1 shows that while the log likelihood index (−2lnL) decreased successively from one to three classes, the Bayesian Information Criterion (BIC) increased for three classes. Thus, the two-class solution, with 68.7 and 31.2% of examinees in Class 1 and Class 2 respectively, was selected for further study. The latent trait means differed significantly between classes ($F_{1,801} = 439.195$, $p < .001$), with Class 1 ($M = 1.143$, $SD = .984$) scoring higher than Class 2 ($M = -.413$, $SD = .865$). Significant racial ethnic differences were observed between the classes $\left(\chi^2_{1,695} = 12.958,\ p < .001\right)$, with 75.0% of Caucasians and 57.3% of African-Americans in Class 1.

Table 1 Mixture Rasch modeling results

Number of classes	Parameters	−2lnL	BIC
Study 1			
1	31	25,472	25,680
2	62	25,146	25,567
3	93	25,044	25,680
Study 2			
1	33	13,222	13,423
2	67	13,068	13,477
3	101	13,001	13,616

Table 2 LLTM weights, standard errors and t value by class

Complexity source	Class 1 ($df = 572$, $\Delta = .820$)			Class 2 ($df = 229$, $\Delta = .809$)		
	Weight	SE	t value	Weight	SE	t value
Unique elements	.1922	.0113	16.95*	.2681	.0185	14.50*
Memory load	.1851	.0049	37.49*	.0926	.0077	12.09*
Integration	.4543	.0454	10.00*	.5502	.0622	8.85*
Distortion	.7434	.0654	11.36*	−.0121	.1054	−.12
Fusion	.3150	.0508	6.20*	.0549	.0723	.76
Intercept	−4.1809	.1018	−41.08*	−2.2618	.1285	−17.61*

*$p < .01$

LLTM was applied within each class to determine the relative impact of the sources of cognitive complexity. While the overall prediction, as indicated by the Δ statistic (Embretson, 1999) shown on Table 2, was strong for both classes, the LLTM weights for cognitive complexity differed. Typically, the strongest predictor is Memory Load; however, the weight for Memory Load was significantly higher in Class 1. Unique Elements was the strongest predictor in Class 2 and two of three perceptual complexity variables were not significant.

Item difficulty also was modeled by the sources of memory load from the five types of relationships. It was found that the number of Figure-Addition relationships was correlated negatively for Class 1 ($r = -.211$) and positively for Class 2 ($r = .216$). Items with Figure-Addition relationships mostly more difficult for Class 2 (see Fig. 3).

Finally, ART trait estimates were correlated with four factors of ASVAB: Verbal, Quantitative, Perceptual Speed and Technical Information. Although significant positive correlations were found with all factors except Perceptual Speed for Class 1, no significant correlations with ASVAB factors were found for Class 2.

Discussion. Two classes of examinees, with varying patterns of item difficulty, were identified on the ART for fluid intelligence. Class 2 was characterized by substantially lower trait levels and lack of significant correlations with other aptitude

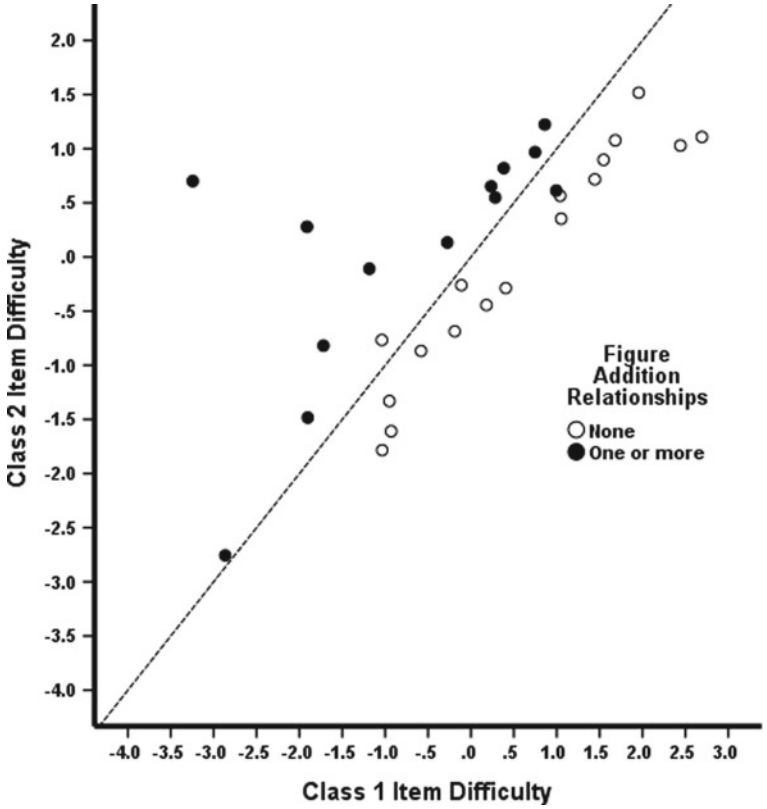

Fig. 3 Item difficulties by class

measures (i.e., ASVAB factors). Further, item difficulty was less predictable for Class 2 from the memory load associated with ART items. An analysis of the relationship types that contribute to memory load indicated that items with Figure-Addition relationships had substantially higher difficulty in Class 2. A possible explanation is that examinees in this class were unfamiliar with the Figure-Addition relationships and applied the much harder Distribution of Two relationship. Figure 4 shows examples of these relationships. Notice that the item on the left requires two Distribution of Two relationships (i.e., changes in the hourglass and house figures), as well as a Constant in a Row (triangles). The item on the right, however, can be solved by either three Figure-Addition (colum 3 is the substraction of column 2 from column 1) or three Distribution of Two relationships.

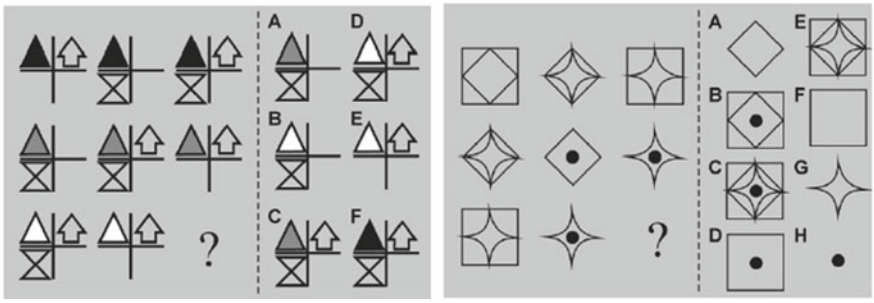

Fig. 4 Two ART items varying in distribution of two relationships

3.3 Study 2

The application of the mixture Rasch model in Study 1 identified a class of examinees with lower scores and different patterns of item difficulty that may be based on unfamiliarity with the possible types of relationships that can occur in ART. In this study, instructions were added to demonstrate each type of relationship.

Method. The examinees were 444 police recruits who were enrolled in basic training in law enforcement. A version of ART with 32 items included extended instructions in which all types of relationships were presented and illustrated. These instructions involved approximately eight additional minutes of testing time. For a sub-sample of examinees, training scores and scores on another test of fluid intelligence were available.

Results. The test was somewhat easy for the sample based on raw scores ($M = 21.459, SD = 4.779$) and latent trait estimates ($M = 1.152, SD = 1.203$). As for Study 1, racial-ethnic comparisons were made between groups with N > 50. The latent trait estimates were significant ($F_{2,406} = 3.099$, $p = .016$, $\eta^2 = .015$). Compared to Caucasians, standardized differences of ($d = .276$) for African Americans and ($d = .075$) for Hispanics were observed.

The mixture Rasch model was applied to determine the number of classes. Table 1 shows that while the log likelihood index (−2lnL) decreased somewhat from one to two classes, the BIC index increased. Thus, the single class model is the preferred solution. Finally, for a subsample of 144 recruits, scores for a six-week course in legal issues for police officers were available. Training scores were correlated more highly with ART ($r = .333$, $p < .001$) than with the *Cattell Culture Fair Intelligence Test* (CCF; $r = .211$, $p = .009$).

Discussion. A single item-solving strategy is supported for ART when administered with extended instructions. That is, a single class was supported with mixture Rasch modeling. Further, the magnitude of the racial-ethnic differences was also substantially smaller in this study. Finally, ART correlated more highly with training than a similar non-verbal intelligence test, which has very short instructions.

4 Summary

The purpose of this chapter was to illustrate how using explanatory IRT models can contribute to the test development process and impact validity. The mixture Rasch model identified two classes of examinees on the ART with different item difficulty orders. The LLTM indicated strong predictability of item performance from cognitive complexity variables; however, the weights varied by class, supporting strategy differences. Items involving a certain type of relationship were relatively more difficult in the lower scoring class. Further, there was an undesirable impact of the second class on the *external relationships* aspect of validity; ART did not correlate with other aptitude tests and racial-ethnic differences were also found. A redesigned ART, that include extended instructions on types of relationships, had a single class, supporting common problem-solving strategies. Further, racial ethnic differences were substantially smaller on the redesigned ART and ART had stronger correlations with achievement than a similar test of fluid intelligence. Thus, two explanatory IRT models were used to inform the *responses processes* aspect of validity for a fluid intelligence test. The redesigned test to optimize responses processes had smaller racial-ethnic differences than the previous ART and more desirable external relationships than the CCF, a similar test of fluid intelligence.

References

American Educational Research Association, American Psychological Association & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of processing in the Raven's Progressive Matrices Test. *Psychological Review, 97,* 404–431.

De Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York: Springer.

Embretson, S. E. (1984). A general multicomponent latent trait model for response processes. *Psychometrika, 49,* 175–186.

Embretson, S. E. (1997). Multicomponent latent trait models. In W. van der Linden & R. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 305–322). New York: Springer.

Embretson, S. E. (1999). Generating items during testing: Psychometric issues and models. *Psychometrika, 64,* 407–433.

Embretson, S. E. (2016). Understanding examinees' responses to items: Implications for measurement. *Educational Measurement: Issues and Practice*, *35*, 6–22.

Embretson, S. E. (2017). An integrative framework for construct validity. In A. Rupp & J. Leighton (Eds.), *The handbook of cognition and assessment* (pp. 102–123). New York: Wiley-Blackwell.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37,* 359–374.

Fischer, G. H., & Ponocny, I. (1995). Extended rating scale and partial credit models for assessing change. In G. H. Fischer & I. W. Molenaar (eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 353– 70). New York: Springer.

Glas, C. A. W., van der Linden, W. J., & Geerlings, H. (2010). Estimation of the parameters in an item cloning model for adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 289–314). New York: Springer.

Janssen, R. (2016). Linear logistic models. In W. van der Linden (Ed.), *Handbook of item response theory: Models, statistics and applications*. New York: Taylor & Francis Inc.

Janssen, R., Schepers, J., & Peres, D. (2004). Models with item and item group predictors. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 189–212). New York: Springer.

Janssen, R., Tuerlinckx, F., Meulders, M., & De Boeck, P. (2000). A hierarchical IRT model for criterion—referenced measurement. *Journal of Educational and Behavioral Statistics, 25,* 285–306.

Linacre, J. M. (1989). *Multi-facet Rasch measurement*. Chicago: MESA Press.

Molenaar, D., & De Boeck, P. (2018). Response mixture modeling: Accounting for heterogeneity in item characteristics across response times. *Psychometrika, 83,* 279–297.

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research, 47,* 667–696.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement, 3,* 271–282.

van der Linden, W. (Ed.). (2016). *Handbook of item response theory: Models, statistics and applications*. New York: Taylor & Francis Inc.

von Davier, M., & Rost, J. (1995). Polytomous mixed Rasch models. In G. Fischer & I. Molenaar (Eds.), *Rasch models: Foundation, recent developments and applications* (pp. 371–379). New York: Springer.

A Taxonomy of Item Response Models in Psychometrika

Seock-Ho Kim, Minho Kwak, Meina Bian, Zachary Feldberg, Travis Henry, Juyeon Lee, Ibrahim Burak Olmez, Yawei Shen, Yanyan Tan, Victoria Tanaka, Jue Wang, Jiajun Xu and Allan S. Cohen

Abstract The main aim of this study is to report on the frequency of which different item response theory models are employed in *Psychometrika* articles. Articles relevant to item response theory modeling in *Psychometrika* for 82 years (1936–2017) are sorted based on the classification framework by Thissen and Steinberg (Item response theory: Parameter estimation techniques. Dekker, New York, 1986). A sorting of the item response theory models used by authors of 367 research and review articles in Volumes 1–82 of *Psychometrika* indicates that the usual unidimensional parametric item response theory models for dichotomous items were employed in 51% of the articles. The usual unidimensional parametric item response theory models for polytomous items were employed in 21% of the articles. The multidimensional item response theory models were employed in 11% of the articles. Familiarity with each of more complicated item response theory models may gradually increase the percentage of accessible articles. Another classification based on recent articles is proposed and discussed. Guiding principles for the taxonomy are also discussed.

Keywords Item response theory · Models · Psychometrika · Rasch model · Taxonomy

1 Introduction

In this study, we report on the frequency of use of item response theory models in *Psychometrika* classified using the taxonomy of Thissen and Steinberg (1986) to answer the following questions: Will knowledge of a few basic item response theory models, such as the Rasch model and the three-parameter logistic model, assist readers in recognizing the modeling component of a high percentage of research articles that are relevant to item response theory modeling in *Psychometrika*? Which additional

S.-H. Kim (✉) · M. Kwak · M. Bian · Z. Feldberg · T. Henry · J. Lee · I. B. Olmez · Y. Shen · Y. Tan · V. Tanaka · J. Wang · J. Xu · A. S. Cohen
University of Georgia, Athens, GA 30602-7143, USA
e-mail: shkim@uga.edu
URL: https://coe.uga.edu/directory/people/shkim

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings in Mathematics & Statistics 265, https://doi.org/10.1007/978-3-030-01310-3_2

item response theory models are used most often and therefore could be added most profitably to the psychometric and educational measurement background of readers? To aid psychometricians, measurement specialists, and applied statisticians who are continuing their own psychometric training, as well as persons designing courses in psychometrics and educational measurement for advanced undergraduate and graduate students, we report on modeling components of the item response theory relevant research and review articles in *Psychometrika* between 1936 and 2017.

In their taxonomy, Thissen and Steinberg (1986) classified item response theory models into four distinct groups based on assumptions and constraints on the parameters: binary models, difference models, divided-by-total models, and left-side-added models. They classified, for example, the two-parameter normal ogive model and the Rasch model as the binary models; Samejima's graded response model in normal ogive and logistic forms as the difference model; Bock's nominal response model and Master's partial credit model as the divide-by-total models; and Birnbaum's three-parameter logistic model as the left-side-added model (see Thissen & Steinberg, 1986, and references therein). In this paper, we present a more refined classification of the item response theory models based on the type of data analyzed.

2 Methods

This study analyzed Volumes 1 through 82 (March 1936 through December 2017) of *Psychometrika* and included all articles identified in the table of contents as Articles, Notes, Comments, Brief Comments, Tables, and Presidential Addresses. Excluded were Errata, Announcements, Abstracts, Book Reviews, Rules, Obituaries, Reports, Minutes, Notices, Constitution, and Lists of Members. For example, the excluded portions included: Volume 1, Issue 2, Pages 61–64 that contained the List of 150 Members of the Psychometric Society; Volume 4, Issue 1, Pages 81–88 that contained the List of 235 Members of the Psychometric Society; and Volume 2, Issue 1, Pages 67–72 that presented the Abstracts of 11 papers to be presented at the District Meeting of the Psychometric Society, The University of Chicago, on Saturday, April 3, 1987.

2.1 Review Process

Initially, 2837 articles were screened and identified from these volumes, and a group of measurement specialists eventually selected 367 articles for detailed review. The 367 articles were selected for their relevance to various models in item response theory. At least two measurement specialists independently reviewed each of the 367 articles for their use of item response theory models and completed a checklist

documenting topics and models. The excluded articles received a second but briefer review for the presence or absence of the use of item response theory models in the procedures and techniques employed in the study. The reviewer read the abstracts, the methods sections, and all tables, and scanned other sections of the articles for the pertinent information. All reviewers were faculty members or graduate students trained in both quantitative methodology and applied statistics.

For the 367 articles receiving detailed review, any discrepancies between the two or more independent reviewers were discussed and resolved. Discrepancies were found initially for some of these articles. Another careful reading of these discrepant articles by the reviewers indicated that nearly all errors involved overlooking the methods section and the procedures and techniques used in the article.

For the 367 articles relevant to item response theory modeling in this study, we first partitioned these papers into theoretical and application types. Due to the characteristic of *Psychometrika* as a leading journal in psychometrics, the articles except for four were classified as theoretical.

2.2 Analysis of Models Used

We determined the frequency of the item response theory models in the 367 journal articles. Articles were sorted based on the classification framework by Thissen and Steinberg (1986). In addition to performing the simple quantification (number and percentage of articles using a method), we assessed how much a reader's acquaintance with additional item response theory models would improve his or her psychometric repertoire. In trying to obtain a definite measure, we were handicapped by the lack of a natural order for learning and applying these models. For the analysis, we chose the order that maximally increased the percentage of articles for which a reader would be acquainted with the item response theory models employed if he or she learned one more item response theory model.

We began this analysis by assuming that there are three major ordered classes of the item response theory models; (1) unidimensional parametric item response theory models for dichotomous items, (2) unidimensional parametric item response theory models for polytomous items, and (3) multidimensional item response theory models. In a sense, the order was thus determined by the complexity of models as well as modeling data gathered. This ordering, though useful, intellectually reasonable, and empirically based, is nevertheless arbitrary. In particular, it ignores the fundamental role of broad psychometric concepts used in the article such as adaptive testing, differential item functioning, equating and linking, parameter estimation techniques, test scoring, and so on in determining the extent of a reader's psychometric understanding. Furthermore, it may not be the best order for learning about the item response theory models.

3 Results

Figure 1 shows the time plots of the number of articles in *Psychometrika* as well as the number of item response theory relevant articles in each volume from 1936 to 2017. The average of the number of articles in each volume was 34.6 and its standard deviation was 8.6. The five number summary was (19, 28.8, 33, 41, 59). There was a steady increasing pattern in terms of the number of articles in each volume. The average of the number of item response theory relevant articles in each volume was 4.2 and its standard deviation was 4.3. The five number summary was (0, 0, 2.5, 8, 17). A rapid increase occurred between the 70's and the 90's for the number of item response theory relevant articles in each volume.

Figure 2 shows the time plot of the proportion of the item response theory relevant articles in each volume from 1936 to 2017. The average of the proportion was .11 and its standard deviation was .11. The five number summary was (0, 0, .07, .21, .53). The proportion was rapidly increased between the 70's and the 90's.

Table 1 presents the number of articles that used different item response theory models by decades from the 1930s (n.b., the 1930s starts from 1936) to the 2010s (n.b., the 2010s are not finished yet). The bottom line contains the total number of unique (i.e., not the column sum) item response theory relevant articles by decades. The far right-hand-side column of Table 1 shows the frequency of item response theory models found in Volumes 1 through 82 of *Psychometrika*. Under the assumptions outlined above, we analyzed the frequencies of the classes of item response theory models employed in the journal articles.

The followings are the model acronyms in Table 1: One-Parameter Logistic (1PL), One-Parameter Normal (1PN), Two-Parameter Logistic (2PL), Two-Parameter Normal (2PN), Nonparametric (NON), Three-Parameter Logistic (3PL), Three-Parameter

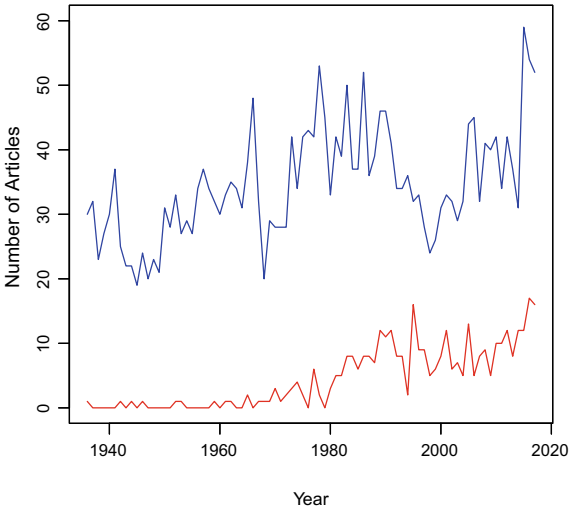

Fig. 1 Time plots of the number of articles in blue and the number of item response theory relevant articles in red

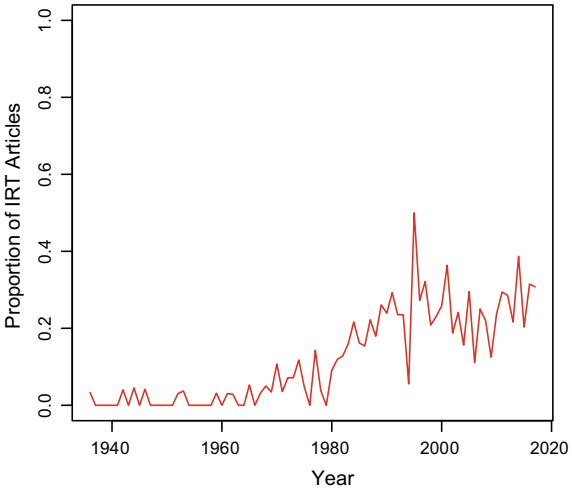

Fig. 2 Time plot of the proportion of item response theory relevant articles

Normal (3PN), Two-Parameter of Choppin (2PC), Four-Parameter Logistic (4PL), Multiple Choice of Samejima (MCS), Multiple Choice of Thissen and Steinberg (MCTS), Multiple Choice (MC), Graded Response (GR), Partical Credit (PC), Rating Scale (RS), Generalized Partial Credit (GPC), Nominal Categories (NC), Binomial Trials (BT), Poisson Counts (POC), Continuation Ration (CR), Linear Logistic Test Model (LLTM), and Multidimensional Item Response Theory (MIRT).

Table 1 shows many articles reviewed relied on some type of unidimensional dichotomous item response theory models. These articles used the Rasch model most frequently by 107 out of 367 articles. The one-parameter logistic model with a common item discrimination parameter was used in 15 articles. The two-parameter logistic model was used by 60 out of 367 articles, and the two-parameter normal ogive model was used by 37 out of 367 articles. The three-parameter logistic model was used quite frequently, that is, 82 out of 367 articles. The polytomous item response theory models are generally used less frequently (25 for the graded response model, 21 for the partial credit model, 10 for the rating scale model, 5 for the generalized partial credit model, and 7 for the nominal categories model).

It can be noticed that the various taxonomic classifications of the item response theory models defined in Table 1 were not frequently employed in the articles reviewed. The impression is that only limited cases of the item response theory models or the combinations of the models have been employed in *Psychometrika*, although this finding does depend on the initial taxonomy of the item response theory models. Articles published recently within about 20 years that used item response theory models were more complicated both mathematically and statistically than other previously published articles in *Psychometrika*. Theoretical research studies based on more complicated item response theory models require a deeper understanding of and more extensive training in psychometrics and applied statistics.

Table 1 Item response theory models from psychometrica articles

Taxonomy type	Model	Period									
		1930s	40s	50s	60s	70s	80s	90s	2000s	10s	Row total
Binary	Rasch					7	33	27	15	25	107
	1PL						6	2	1	6	15
	1PN										
	2PL			1	2		11	16	9	21	60
	2PN	1	3	3	5	2	5	4	8	6	37
	NON						1	7	9	6	23
Left-side-added (LSA)	3PL					6	20	19	14	23	82
	3PN						1				1
	2PC										
	4PL								1	1	2
LSA-DBT	MCS										
	Model 6										
	MCTS						1				1
	MC								1	1	2
Difference	GR				2	4	2	9	2	6	25
Divided-by-total (DBT)	PC						5	7	3	6	21
	RS					1	5	1	2	2	10
	GPC							4	1		5
	NC					2	1	1	3		7
	BT										
	POC							2	1		3
	CR					2			1		3
Extension	LLTM						3	2	6	1	12
	MIRT						1	6	10	17	34
	Testlet							1		1	2
	Multilevel								3	3	6
	Other						5	4	9	5	23
Unique articles	No.	1	3	3	7	23	70	85	78	97	367

A sorting of the item response theory models used by authors of the 367 articles in *Psychometrika* indicates that a reader who is familiar with the usual unidimensional parametric item response theory models for dichotomous items (e.g., the Rasch model, the one-parameter logistic model, the two-parameter logistic or normal ogive model, and the three-parameter logistic or normal ogive model) may have potential access to 186 out of 367 articles (51%). Note that the number 186 was not obtained from Table 1 but based on the separate counting of the articles. Note also that the numbers in Table 1 are not mutually exclusive because, for example, an article might employ two or more different item response theory models together. It should also be noted that the accessibility here implies the recognition of the model used in the article instead of comprehension of the entire contents of the article. Because the unidimensional parametric item response theory models for polytomous items (e.g., the graded response model, the partial credit model, the rating scale model, the nominal categories model, and the generalized partial credit model) were employed in 79 out of 367 articles, a reader who is familiar with the two classes of the unidimensional item response theory models may have potential access to cumulatively 72% of the journal articles. Familiarity with each of the more complicated item response theory models may gradually increase the percentage of accessible articles. If one knew the multidimensional item response theory models in addition to the unidimensional item response theory models, one would access 38 articles, or 83 cumulative per cent of the number of articles reviewed. However, more complicated models (e.g., nonparametric models, testlet models, mixture models, multilevel models, etc.) were concurrently used in the psychometric research journal articles together with the usual parametric models for the dichotomous and polytomous items. Hence, 64 out of 367 (17%) of the articles cannot be fully accessible in terms of item response theory if a reader is familiar with only these parametric models.

Although some classifications were obviously quite narrowly defined, others such as multidimensional item response theory models and nonparametric models were not. Furthermore, these latter models, though cited infrequently in the articles, may be more frequently used in other application fields and may become more common in future psychometric research.

The selected articles relevant to item response theory modeling in Table 1 were sorted based on the classification framework by Thissen and Steinberg (1986). Another recent classification based on Van der Linden (2016a), however, can be used, and a more refined subclassification (e.g., Nering & Ostini, 2010) can also be considered. Note that articles may be further sorted by the parameter estimation methods (e.g., Baker & Kim, 2004; De Ayala, 2009) as well as the computer programs used to implement the estimation methods (e.g., Hambleton, Swaminathan, & Roger, 1991, pp. 159–160; Van der Linden, 2016b).

Psychometric researchers interested in continuing their own training in methodology should find the frequencies of various item response theory models presented in Table 1 helpful in identifying the knowledge of which item response theory models they should be aware. This paper reviews item response theory models with the perspective of a general reader, and no attempt has been made to identify a hierar-

chical structure of the item response theory models, which may vary for researchers in different psychometric research areas within different specialties.

4 Discussion

Except for general item response theory review articles in *Psychometrika*, not many item response theory models were used simultaneously in each research article. As noted by Popham (1993) and Bock (1997) there are several unexpected consequences of using item response theory models in the analysis of testing data. Only a limited number of item response theory experts can fully understand what is happening, for example, in the process of test calibration. Also, there are many different directions of the development of item response theory so that even experts in the item response theory field may not be able to comprehend the full scope of the theory and applications of item response theory. It is unfortunate that the item response theory models and item response theory itself are too difficult to understand for scholars with only limited statistical and mathematical training. Nevertheless, item response theory does occupy and may continue to occupy major portions of lively and productive future development in psychometric research.

Understanding some of the item response theory relevant articles in *Psychometrika* definitely requires more than the familiarity of the item response theory models. For example, training in modern Bayesian statistics for which the Bayesian posterior approximation methods with data augmentation techniques are taught is needed for reading several articles. Note that the normal ogive models were frequently employed in data augmentation techniques by some authors who are themselves prepared for understanding more advanced research articles.

It should be noted that the numerical measure of ability or proficiency is the ultimate, eventual entity that is pursued in item response theory modeling. In other applications, the item parameters are something needed assuming that persons are randomly sampled from a population. In item response theory with such a sampling concept, the item parameters are the structural parameters while the person parameters are incidental parameters. The concept of invariance of ability with regard to the sets of item parameters (i.e., persons ability can be measured with different sets of items) as well as invariance of item characteristics with regard to the groups of persons (i.e., item characteristics can be obtained with different groups of persons) are crucial in item response theory. Many investigations of structural parameters such as measurement invariance or differential item functioning studies are studies of structural parameters. Note that measurement invariance is a preliminary to studying invariant person measures, and as such, needs to be seen as a process within measurement validation (Kane, 2006). Hence, item response theory models and the required parameters to estimate ought to be scrutinized in conjunction with specific application areas.

In the field of educational assessment, items can be in the forms of both dichotomously-scored and polytomously-scored. In most large scale assessment pro-

grams (e.g., National Assessment of Educational Progress, Trends in International Mathematics and Science Study) a combination of the three-parameter logistic model and the generalized partial credit model is used to calibrate item response data. In the analysis of instruments with mixed item types, there are special combinations of dichotomous and polytomous item response theory models to be used (e.g., Rasch and PC; 2PL and GR). So, there are natural combinations of item response theory models for mixed item types.

This study may be helpful to people designing and teaching courses in psychometric methods for advance undergraduate and graduate students and other psychometricians or measurement specialists using various item response theory models. But one should keep in mind that any professional specialization in psychometric research may influence understanding with regard to the relative importance of the various item response theory models.

The purpose of writing for some journal articles that are relevant to item response theory in psychometric research might not be to disseminate the findings of the studies to more general psychometric researchers. The authors might have tried to demonstrate their capabilities to invent novel models, to create new ideas, and to explore challenging areas of psychometrics. Consequently, there are a plethora of item response theory models invented recently.

We have identified the various models in item response theory that have been used by psychometricians in *Psychometrika* articles and that are thus very much likely to be used by future authors in psychometric research. Note that the latter point may not be the case because some articles used the most esoteric item response theory models together with complicated computational techniques. The appropriate training of psychometric researchers in the use of item response theory models seems to be an important consideration. Such an issue should be addressed by the leading scholars who are responsible for training future psychometric researchers. More in depth evaluation of the articles and more thorough review would be helpful.

It can be noted that item response theory models presented in Table 1 already contained additional models than those (e.g., 4PL, MC, CR, LLTM, MIRT, Testlet, and Multilevel) in Thissen and Steinberg (1986). There are many different item formats so we may classify item response theory models in terms of the item response data and additional variables required for the modeling. If we denote the original item response data for multiple-choice items as U, then item response theory models for multiple-choice items can be used to estimate model parameters. When we denote the keyed or scored data to be R and further denote dichotomously scored data to be D, then we may use the Rasch model and other item response theory models for dichotomously scored items (e.g., 1PL, . . . , 4PL). If R can be further specified with the types of polytomously scored items that is denoted by P, then we may use item response theory models for polytomous items. Here, the set of item parameters can be denoted by ξ and the set of ability parameters can be denoted by θ. If we allow additional dimensionality to the item and ability parameters, then we may have multidimensional item response theory models. In the above context, if there exist a latent group hyperparameter τ and both ability and item parameters are characterized