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

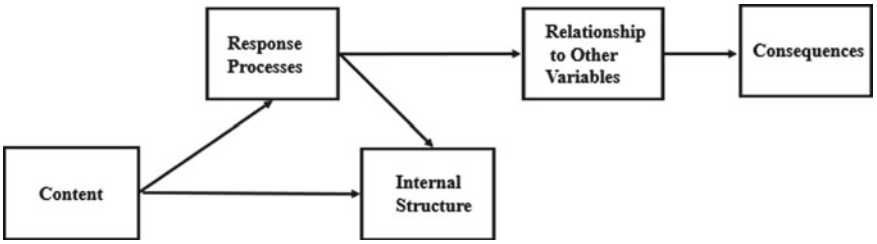

Fig. 1 Unified framework for validity

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.

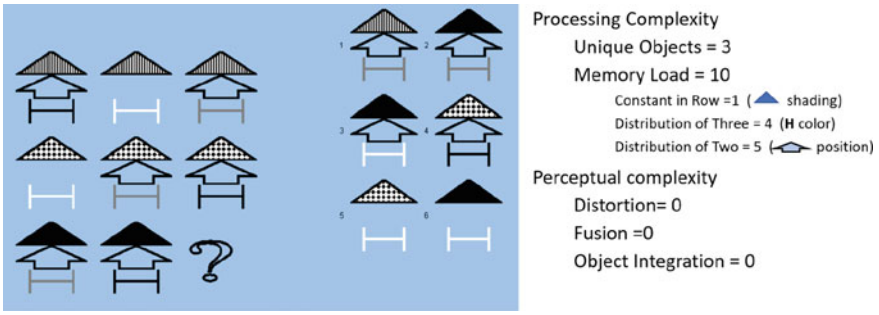

Fig. 2 Example of an ART item

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

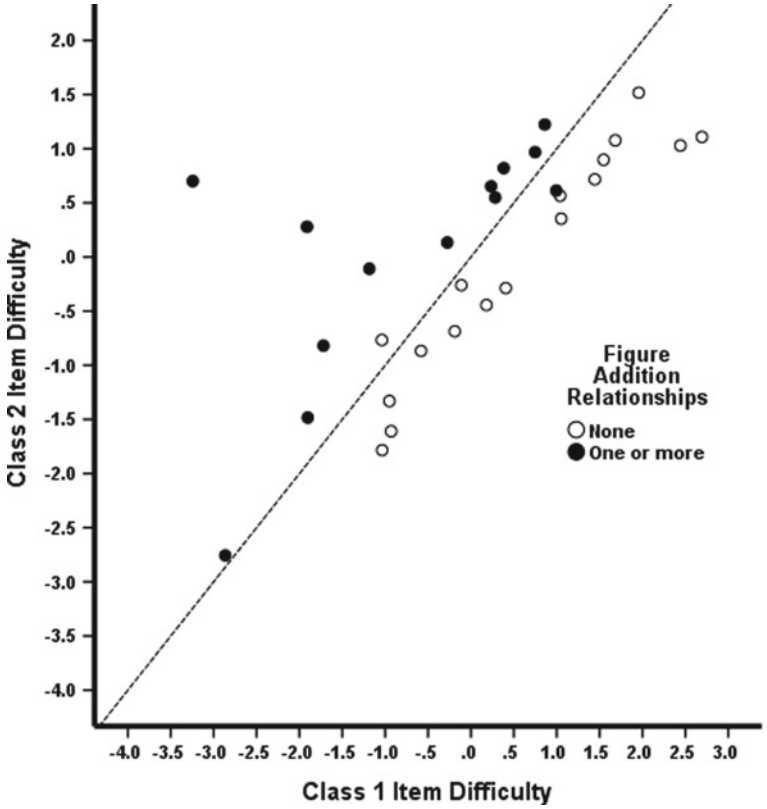

Fig. 3 Item difficulties by class

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.

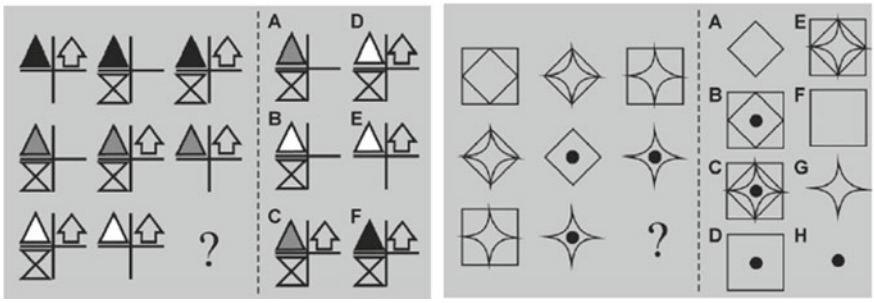

Fig. 4 Two ART items varying in distribution of two relationships

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.