8-1998

Modeling Item-Level Data with Item Response Theory

Michael John Zickar

Bowling Green State University, mzickar@bgsu.edu

Follow this and additional works at: https://scholarworks.bgsu.edu/psych_pub

Part of the [Psychology Commons](#)

How does access to this work benefit you? Let us know!

Repository Citation

Zickar, Michael John, "Modeling Item-Level Data with Item Response Theory" (1998). *Psychology Faculty Publications*. 4.

https://scholarworks.bgsu.edu/psych_pub/4

This Article is brought to you for free and open access by the Psychology at ScholarWorks@BGSU. It has been accepted for inclusion in Psychology Faculty Publications by an authorized administrator of ScholarWorks@BGSU.

Modeling Item-Level Data With Item Response Theory

Michael J. Zickar¹

Department of Psychology, Bowling Green State University, Bowling Green, Ohio

Psychology can be separated into two camps, a substantive camp that is primarily interested in understanding important aspects of human behavior and thought and a methodological camp that is primarily interested in developing tools that will be used by the substantive researchers to answer difficult questions. Ideally, the two camps should have much cross-fertilization, and the lines between them should be blurred. However, complex theories are often constructed before appropriate methods are developed for testing the hypotheses. Additionally, quantitative psychologists often develop techniques that might be difficult and impractical to apply initially to real-data problems. These novel techniques need to be refined through robustness studies and analytical work before they can be used with real-world data. A psychometric framework called item response theory (IRT) has undergone such prudent testing and re-

finement and is now ready to be implemented into the mainstream of psychological research and practice.

IRT has already had a major impact on educational testing through its impact on computerized adaptive testing (CAT). The precision of IRT-based item statistics allows computerized testing programs to choose items that provide maximum psychometric information for an individual examinee. This process allows adaptive tests to maintain measurement precision similar to that of conventionally administered tests even though fewer items are administered. In the 1990s, Educational Testing Service (ETS) implemented a CAT version of the Graduate Record Examination (GRE). ETS plans to phase out conventional paper-and-pencil administration of the GRE General test and administer only CAT versions by fall 1999. The success of adaptive testing would not be possible without development of IRT. In the future, it is likely that IRT will yield progress not only by improving measurement technologies, but also by making contributions in substantive areas, such as decision-making theory.

This article begins by explaining two fundamental concepts of IRT, the item response function and item information. Next, the flexibility of IRT is highlighted to demonstrate the types of data that can be modeled. Finally, the present and future impact of IRT on both practical and theoretical issues in psychology is discussed.

TWO BASIC CONCEPTS OF IRT

Psychologists typically use many of the techniques developed through classical test theory (CTT) to evaluate tests and scales used in research and practice. Concepts such as reliability, item-total correlations, and the Spearman-Brown prophecy formula are all based on the CTT model, which posits that the observed scale score is a function of a true score and an error term. Many limitations of this framework have been noted. First, the CTT model focuses on scale-level measurement. The true score is explicitly linked to a particular instrument. Therefore, an individual's true score for one mathematical reasoning test is different from his or her true score for another mathematical reasoning test. Second, CTT does not allow for much precision in testing specific hypotheses about the measurement properties of scales. In CTT, each scale has a reliability that is used to characterize the measurement precision of the entire test. This concept fails to recognize that tests have differential capabilities in discriminating among different levels of examinees' abilities. For example, a mathematics test composed primarily of calculus items will be able to differentiate individuals high in math ability from those with average or below-average skills; this test, however, would not provide much differentiation between those of low ability and those of average ability. To appreciate the power of IRT and some of its advantages over CTT, it is necessary to discuss two basic concepts.²

Item Response Function

IRT relates characteristics of items and characteristics of indi-

Recommended Reading

- Drasgow, F., & Hulin, C.L. (1990). Item response theory. In M.D. Dunnette & L.M. Hough (Eds.), *Handbook of industrial and organizational psychology*, Vol. 1 (pp. 577-636). Palo Alto, CA: Consulting Psychologists Press.
- Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: SAGE.
- van der Linden, W.J., & Hambleton, R.K. (Eds.). (1997). *Handbook of modern item response theory*. New York: Springer-Verlag.

viduals to the probability of affirming, endorsing, or correctly answering individual items. The cornerstone of IRT is the *item response function* (IRF). This function is a nonlinear regression of the probability of affirming item i on a latent trait, θ , which represents the characteristic measured by the scale items (e.g., mathematical ability, extroversion, job satisfaction). Figure 1 presents a graphic representation of three IRFs.

There are many different forms of this regression line. For dichotomously scored items (e.g., right vs. wrong or true vs. false), the two-parameter logistic model (2PL) and the three-parameter logistic model (3PL) are commonly used. The formula for the 3PL model is

$$P(u_i = 1|\theta) = c_i + (1 - c_i) \frac{1}{1 + e^{-1.7a_i(\theta - b_i)'}}$$

where the probability that a person with a latent trait, θ , affirms an item i (i.e., $u_i = 1$) is a function of three parameters: a_i , a discrimination parameter that determines the

slope of the IRF; b_i , a location parameter that determines the area of the θ continuum in which the IRF is most steep; and c_i , a pseudoguessing parameter that determines the probability that a respondent with an extremely low θ will endorse the item. The probability of affirming items with large a parameters varies sharply as a function of θ , whereas the probability of affirming items with low a parameters varies weakly as a function of θ . Items with low a parameters are generally considered poor items. Items with large positive b parameters will be endorsed only by respondents with large positive θ s, whereas items with large negative b parameters will be endorsed by everyone except people with the most extreme negative θ s. The c parameter introduces a nonzero lower asymptote to the IRF so that respondents with large negative θ s will have a nonzero probability of affirming the item; this nonzero asymptote may result from guessing or other processes. The 2PL formula is a submodel of the 3PL for-

mula and can be obtained by setting the c parameter to zero. This model has the implicit assumption that people with the lowest θ values will have a zero probability of affirming the item. An even simpler model, the Rasch model, is obtained by setting the a parameter to be constant across all items. Each of these models assumes that each item is measuring only one θ dimension.

The first hypothetical item presented in Figure 1 has $a = 1.0$, which is an average level of discrimination; $b = 0.0$, which is an average difficulty; and $c = .25$, which suggests that even individuals with extremely low ability have a 25% chance of answering the item correctly. Item 2 has the same discrimination and difficulty as Item 1; however, $c = 0.0$, suggesting that there is no guessing occurring with this item. As can be seen in Figure 1, the lower asymptote of the IRF for Item 2 is at 0.0. Finally, Item 3 is of lower discrimination ($a = 0.60$) and of lower difficulty ($b = -1.5$) than the previous two items. This item has $c = 0.0$; if the θ axis were extended further, the IRF would eventually reach 0.0. By plotting the IRFs, researchers can compare the functioning of items, determine the extent of guessing, and determine the range of θ in which an item is most discriminating.

Information

Another key IRT concept is *information*, which is used to quantify measurement precision. The information value for each item is computed on the basis of the IRF. The formula for item information is

$$I_i(\theta) = \frac{[P(u_i = 1|\theta)]^2}{[P(u_i = 1|\theta) * [1 - P(u_i = 1|\theta)]]},$$

where $P(u_i = 1|\theta)$ is the first derivative (i.e., slope) of the IRF at a particular value of θ .

Fig. 1. Item response functions for three hypothetical items.

The standard error of θ can be computed directly from the item information function as follows:

$$SE(\theta) = \frac{1}{\sqrt{I(\theta)}}$$

As can be seen in this equation, the standard error of measurement as conceptualized in IRT is conditioned on the level of θ . Unlike CTT, IRT allows measurement precision to vary across different regions of the measurement scale.

Figure 2 presents the information function for each of the three previously presented items. Several important observations can be made about the information function. First, information is maximized near where θ equals the value of the b parameter of each item. The functions for Items 1 and 2 have more information near the middle of the θ distribution than does the function for Item 3, which has its information maximized in the negative range of the θ continuum. This difference illustrates

an interesting point: Items differ in their usefulness depending on the purpose. Item 3, which has a much lower discrimination parameter than Items 1 and 2, might be preferred if the purpose is to estimate the θ of a person with low ability because that item has a higher amount of information at the low end of the θ continuum. A second important observation, that guessing reduces information, can be seen by examining the difference between information functions for Item 1 and Item 2. Item 2 has greater information at all regions of the θ continuum, and especially at the lower regions of the θ continuum, where guessing will be prevalent among respondents. This decrease in the amount of information highlights the negative effects that guessing has on measurement precision.

Computer adaptive tests work by administering items with large amounts of information near the region where θ is most likely to be. The ability estimate is revised after

each item response, and then a computer algorithm selects the next item to present on the basis of the information level of items at the revised ability estimate. By choosing only items with large amounts of information, adaptive tests can maintain measurement precision at the levels of conventional tests even though fewer items are administered.

EXPANDING THE DOMAIN

Two unfounded criticisms of IRT are that IRT models are capable of modeling only items that measure cognitive ability and that IRT is incapable of modeling items that are not dichotomously scored. IRT models have been developed primarily with cognitive-ability items having clear right and wrong answers. These types of items lend themselves well to dichotomous scoring. However, dichotomous IRT models can be used for items that have two options, even though neither option is considered more correct than the other. For example, personality items that ask the respondent whether a particular description "describes yourself" or "does not describe yourself" can be modeled with dichotomous models. Dichotomous IRT models have been used to model items in diverse fields, such as personality (Zickar & Drasgow, 1996), drug-use screening (Kirisici, Tarter, & Hsu, 1994), and attitudes toward work (Roznowski, 1989).

In recent years, much of the basic research in IRT has focused on developing models for more complicated item types, such as items with more than two response options (i.e., polytomous items) and items that measure more than one dimension. Roznowski (1989), analyzing the Job Descriptive Index (JDI), collapsed a three-option re-

Fig. 2. Item information functions for three hypothetical items. See Figure 1 for the item parameters.

sponse format into a dichotomous response scale because polytomous IRT was relatively unknown and untested at that time. She predicted that "algorithms that allow . . . analysis of [polytomous] JDI data using such methods likely will be forthcoming" (p. 813). She was correct.

Polytomous Models

Polytomous models are appropriate when items are not dichotomously scored and when the attractiveness of options differs for respondents of differing θ levels. The main difference between polytomous and dichotomous models is that IRFs are replaced by option response functions (ORFs). An ORF explicitly relates θ to the probability of choosing a particular option instead of the probability of answering an item correctly.

There are two classes of polytomous models. Models in the first class, denoted graded models, assume an a priori ordering of options in terms of valence, so that Option 1 has a more negative valence than Option 2, and Option 3 has a more positive valence than Options 1 and 2. This model is appropriate with Likert-type scales (e.g., when response options are "disagree," "neither agree nor disagree," "agree"), which are prevalent in personality and attitude measurement. An example of a graded model is Samejima's Graded Response Model (Samejima, 1969).

The other class of polytomous models, denoted nominal models, assumes no a priori ordering of response options within an item. An example of a nominal model is Bock's Nominal Model (Bock, 1972). A nominal model might be most appropriate for cognitive-ability items for which the incorrect options may imply different degrees of "wrongness," but often

the test maker cannot identify the degree of wrongness without looking at the data in a pilot sample. For example, on a vocabulary test, one of the wrong answers may be more similar in meaning to the correct answer than other wrong answers are. For most personality tests, nominal models are probably inappropriate.

Polytomous models provide two advantages to researchers. First, these models can handle more flexible item types. Often (e.g., Roznowski, 1989) polytomous response scales are collapsed so that dichotomous models can be used. This is bad practice because the model used to analyze the data does not match the original response scale. Second, polytomous models can extract more information than dichotomous models used on the same items. This increased information helps improve the precision of θ estimates. Given these two advantages, polytomous IRT has been an important addition to the IRT toolbox.

Multidimensional IRT

The previously mentioned IRT models all assume that the latent trait, θ , is the only individual determinant of item responses. If this prescription were to be taken literally, the models could be used to analyze only scales with strict unidimensionality. However, Drasgow and Parsons (1983) demonstrated that deviations from strict unidimensionality will not destroy the fidelity of these models as long as an appropriate factor analysis demonstrates that the first factor has a much larger eigenvalue than secondary factors. Unfortunately, many psychological constructs, such as personality traits, may have higher levels of multidimensionality than the levels Drasgow and Parsons found acceptable. For these multidimensional constructs,

IRT models that assume unidimensionality will not fit the data.

Recent work has focused on developing multidimensional IRT models (see Reckase, 1997). These techniques can be useful for modeling complex items, such as mathematical story problems, which require both quantitative and verbal skills. However, these techniques still require more work before they will be integrated into substantive research areas frequently. Future work with multidimensional IRT will focus on developing polytomous IRT models and also on making estimation programs more user-friendly. This work is important because it will help break through another limitation.

IRT-BASED TOOLS

Some of the most important consequences of IRT are the psychometric tools that rely on its basic concepts. These tools include adaptive testing and appropriateness measurement, among others.

Adaptive Testing

Adaptive testing presents a challenge for psychometricians because adaptive tests must be constructed to maximize the precision of ability estimates for each test taker. CTT methods are not very practical because the item parameters are linked to the sample for which the parameters are calibrated; IRT parameters are sample-invariant. Item parameters estimated from one sample can be applied to examinees who were not in the calibration sample. Another advantage is that item and ability parameters are on the same scale; this feature allows for relatively simple algorithms to be used to select items that are best suited for each individual. IRT models, along with the

development of low-cost desktop computers, have made adaptive testing practical. Adaptive testing is now used extensively by ETS, the U.S. Armed Services, and other licensure testing services. The popularity of adaptive testing in such high-volume, high-stakes testing programs is a testament to the confidence psychometricians have in IRT.

Appropriateness Measurement

Appropriateness measurement attempts to identify individuals who do not fit the existing model for responding to items (Levine & Rubin, 1979). This technique can be used to identify people who are unmotivated, cheating, faking, or otherwise answering in a way unlike most other respondents. The logic of appropriateness measurement is that a model is estimated on a group of respondents who can be considered to be normal or regular. Response patterns are checked against this normative pattern. Individuals whose patterns do not conform to this model are checked further. Additional investigation may reveal that the individual cheated or did not take the test seriously.

FUTURE DIRECTIONS

There are several directions that I predict future IRT research will follow. On the technical side, more sophisticated models will be developed to model more complex item types. Some examples of this trend can be seen in nonparametric models (e.g., Levine & Tsien, 1997) that provide more flexibility than the models discussed in this article, models that incorporate speed of responding (Roskam, 1997), and multidimensional models for polytomous data (Kelderman, 1997).

These models will provide researchers even more modeling options.

Another important methodological research topic that needs more attention is evaluation of goodness of fit. IRT models, like structural equation models, need to be evaluated to see if the models fit the data. In the structural equations literature, there are many alternative goodness-of-fit indices, which evaluate model fit with different standards. In IRT, few options for evaluating model fit exist. Issues such as model modification indices and influence statistics have yet to be tackled in IRT.

The real contributions of IRT to the field of psychology are yet to be realized; the current challenge is to use IRT models and related tools to answer substantive problems. Thissen and Steinberg (1988) demonstrated that IRT models can be used in conjunction with experimental manipulations to test whether such manipulations alter the way respondents answer items. In an extension of this methodology, Highhouse and I used the 2PL model to model responses to risky-choice scenarios under a positive or negative framing condition (i.e., when positive or negative consequences of a decision, respectively, were emphasized; Zickar & Highhouse, 1998). Our results supported the traditional finding that respondents in negative frames are more likely to choose risky alternatives than are respondents in positive frames; however, when plotting IRFs for individual items administered under both positive and negative frames, we identified that the effect of framing differed across scenario types. The classic Asian Disease problem (Tversky & Kahneman, 1981), which is the most frequently used risky-choice scenario, had differences in discrimination and location parameters across experimental condi-

tions, whereas other scenarios had differences in location parameters only across conditions. The detailed description of the IRFs allowed for fine distinctions between items to be noted. This IRT analysis suggested that the concept of framing needs to be reevaluated so that these differences can be explained.

More than 10 years ago, Roskam (1985) stated that "psychometric modeling has been and still is primarily developed as a technology for measurement, with only loose connections to substantive theorizing" (p. 9). With some exceptions, Roskam's statement still holds true today. With the increased accessibility of IRT estimation programs and the continued dissemination of IRT research in substantive journals, this important methodological technology should help advance the state of psychological knowledge.

Notes

1. Address correspondence to Michael J. Zickar, Department of Psychology, Bowling Green State University, Bowling Green, OH 43403; e-mail: mzickar@bgsu.edu.

2. For more information about the differences between CTT and IRT, the reader is referred to Hambleton and Jones (1993).

References

- Bock, R.D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37, 29-51.
- Drasgow, F., & Parsons, C.K. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement*, 7, 189-199.
- Hambleton, R.K., & Jones, R.W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, 12, 38-47.
- Kelderman, H. (1997). Loglinear multidimensional item response model for polytomously scored items. In W.J. van der Linden & R.K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 287-304). New York: Springer-Verlag.
- Kirisci, L., Tarter, R.E., & Hsu, T. (1994). Fitting a two-parameter logistic item response model to

- clarify the psychometric properties of the drug use screening inventory for adolescent alcohol and drug abusers. *Alcoholism: Clinical and Experimental Research*, 18, 1335-1341.
- Levine, M.V., & Rubin, D.B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics*, 4, 269-290.
- Levine, M.V., & Tsien, S. (1997). A geometric approach to two dimensional measurement. In A.J. Marley (Ed.), *Choice, decision, and measurement* (pp. 207-223). Mahwah, NJ: Erlbaum.
- Reckase, M.D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement*, 21, 25-36.
- Roskam, E.E. (1985). Current issues in item response theory. In E.E. Roskam (Ed.), *Measurement and personality assessment* (pp. 3-19). Amsterdam: Elsevier Science.
- Roskam, E.E. (1997). Models for speed and time-limit tests. In W.J. van der Linden & R.K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 187-208). New York: Springer-Verlag.
- Roznowski, M. (1989). An examination of the measurement properties of the Job Descriptive Index with experimental items. *Journal of Applied Psychology*, 74, 805-814.
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, 34(Suppl. 17).
- Thissen, D., & Steinberg, L. (1988). Data analysis using item response theory. *Psychological Bulletin*, 104, 385-395.
- Tversky, A., & Kahneman, D. (1981). The framing of decisions and the rationality of choice. *Science*, 221, 453-458.
- Zickar, M.J., & Drasgow, F. (1996). Detecting faking on a personality instrument using appropriateness measurement. *Applied Psychological Measurement*, 20, 71-87.
- Zickar, M.J., & Highhouse, S. (1998). Looking closer at the effects of framing on risky choice: An item response theory analysis. *Organizational Behavior and Human Decision Processes*, 75, 75-91.

The Social-Cognitive Model of Transference: Experiencing Past Relationships in the Present

Susan M. Andersen and Michele S. Berk¹

Department of Psychology, New York University, New York, New York

Personal experience, as well as psychological theory and research, suggests that relationships with significant individuals from one's past may have a profound impact on present-day relationships. The notion that aspects of past relationships may reemerge in later social relations also forms the basis of the clinical concept of *transference* (Freud, 1912/1958; Sullivan, 1953), which involves old issues in past relationships emerging in new relations, especially in analysis. Transference in everyday life is the focus of our research, even though historically, transference has been examined mainly theoretically and as it pertains to psychotherapy (e.g., Ehrenreich, 1989). Despite its potential importance to social rela-

tions in daily life, until recently, little empirical work of any kind has examined transference (although see Luborsky & Crits-Christoph, 1990).

In our work, we have developed a *social-cognitive* model of transference in everyday social relations (Andersen & Glassman, 1996; Andersen, Reznik, & Chen, 1997; Chen & Andersen, in press; for related models, see Singer, 1988; Wachtel, 1981; Westen, 1988). We have shown that mental representations of significant others are stored in memory, and that the fundamental processes underlying transference are the activation and application of these representations to new people. Such activation and application occur particularly when the new person resembles the significant other. This research provides the first experimental demonstrations of the transference concept and is relevant to a variety of related literatures, ranging from those dealing with relational schemas (Baldwin, 1992; Bugental, 1992) and attach-

ment theory (e.g., Bowlby, 1969; Collins & Read, 1990; Hazan & Shaver, 1987), to those concerned with the self (e.g., Aron, Aron, Tudor, & Nelson, 1991), close relationships (Berscheid, 1994; Murray & Holmes, 1993), and basic processes in social cognition (Higgins, 1996; Higgins & King, 1981).

In this article, we provide an overview of this research. To begin, we describe the basic tenets of the model, highlighting its social-cognitive and clinical origins, and also outline our experimental paradigm. We then summarize the experimental research supporting the model, which has demonstrated transference as measured by inference and memory derived from significant-other representations, as well as by representation-derived evaluation. We also review research that shows the pervasive impact of transference on interpersonal relations, summarizing findings involving affect, motivation, expectancies, interpersonal roles, and self-definition.

Recommended Reading

- Andersen, S.M., & Berk, M.S. (1998). (See References)
- Andersen, S.M., & Glassman, N.S. (1996). (See References)
- Andersen, S.M., Reznik, I., & Chen, S. (1997). (See References)

THE SOCIAL-COGNITIVE MODEL OF TRANSFERENCE

Basic Assumptions

Research suggests that the activation and use of significant-other