ELSEVIER

Journal of Business Research 57 (2004) 184–208

JOURNAL OF
BUSINESS
RESEARCH

Tackling measurement problems with Item Response Theory: Principles, characteristics, and assessment, with an illustrative example

Jagdip Singh*

Department of Marketing and Policy Studies, Weatherhead School of Management, Case Western Reserve University, Cleveland, OH 44106, USA

Abstract

Interest in measurement issues remains unabated, as evidenced by published research in the area of reliability, validity, and, in particular, scale development. At the same time, psychometricians have continued to generate alternative measurement approaches and models at an explosive pace. Surprisingly, these alternative measurement approaches have been slow to diffuse into the marketing literature despite marketers' inherent interest in measurement issues. This paper discusses one such measurement approach, item response theory (IRT), which can potentially address critical measurement concerns. My focus is on identifying basic principles and key characteristics and on providing an assessment for applied researchers. Toward this end, an empirical example of role conflict (RC) and role ambiguity (RA) concepts is included to illustrate IRT principles and amplify the theory's relevance to resolving measurement dilemmas. In addition, I provide a comparison with the current paradigm of measurement—classical test theory (CTT)—to afford a balanced appreciation of the payoffs of adopting the IRT approach.

Keywords: Measurement; Reliability; Validity; IRT; Role conflict; Role ambiguity

> The link between observation and formulation is one of the most difficult and crucial to scientific enterprises.
> Greer (1969, p. 160).

> Unfortunately, a not uncommon phenomenon in the literature today is a 'study' in which the authors sent a nonpretested, nonscaled questionnaire to a convenience sample of uncertain nature in which little or no thought was given to the reliability of the measurement or the meaningfulness of responses.
> A respondent quoted in Campbell et al. (1982, p. 61).

1. Introduction

Consider the following measurement problems facing a marketing researcher today:

• In developing items for a new measurement scale, the researcher faces a choice between selecting items that are either (a) similar to each other and thus maximize reliability (or fidelity) or (b) different from each other, covering the focal construct broadly and thus maximize validity (or bandwidth). The researcher is aware that a tradeoff is involved since, in their metaanalysis, Churchill and Peter (1984, p. 370) observed that maximizing reliability tended to favor selection of "items (which) were so similar (to each other) that they *underidentify* constructs."

• In deciding the direction of wording for a unidimensional set of scale items, the researcher can either (a) have all items worded in the same direction or (b) split the set with half the items worded in the positive direction and the remaining items worded in the negative direction. However, the researcher is aware that, although the latter approach is a better measurement practice, it is also likely to undermine the unidimensionality of items.

• In developing a "short form" of a unidimensional scale to reduce respondent burden, the researcher decides to select the three items with the highest factor loadings. However, the researcher is unclear as to how this will affect and/or compromise construct validity. Neither is the researcher aware of any other reasonable criteria for developing short forms of established scales.

• In working with an important construct that was developed, say 30 years ago, the researcher is not sure whether to (a) use just the original items or (b) modify and/or enhance the original set by including additional items to tap current reflections of the underlying construct that were not anticipated by original developers. Neither is the re-

* Tel.: +1-216-368-4270; fax: +1-216-368-4785.
E-mail address: jxs16@po.cwru.edu (J. Singh).

0148-2963/$ – see front matter © 2003 Elsevier Science Inc. All rights reserved.
doi:10.1016/S0148-2963(01)00302-2

searcher sure of the best approach to judging the adequacy of the original items.

Not unlike the market environment, the field of measurement theory and method has witnessed exciting, unforeseen changes in the last decade. Interest in understanding measurement—the link between observed and latent trait phenomena (what Greer, quoted above, called "formulation")—appears to be intensifying. In 1993, the first *Handbook of Marketing Scales* was published by Sage Publications in cooperation with the Association of Consumer Research, and shortly thereafter, the American Marketing Association came out with its own handbook. Both handbooks have been recently updated and expanded, indicating that "new" scales continue to be developed at a record pace in the social sciences in general and in marketing in particular. For instance, between 1990 and 1998, no fewer than 87 new scales were published in the marketing literature alone (Bearden and Netemeyer, 1999). This intensity is possibly sparked by growing discomfort with available measures and the lack of attention to measurement issues in past research, as the anonymous respondent suggests in the headnote. Contemporary views of marketing research increasingly resonate with Schwab's (1980, p. 34) admonition that theoretical progress has "suffered because investigators have not accorded measurement the same deference as substantive theory (and) as a consequence, substantive conclusions have been generated that may not be warranted." Thus, outstanding measurement problems such as those noted above are gaining attention, and effort is underway to seek appropriate solutions and meaningful insights, as evidenced by this special issue on measurement.

How can a marketing researcher address these measurement problems? Most measurement approaches in marketing rest on classical test theory (CTT; Lord and Novick, 1968). CTT principles are evident in measurement methods ranging from reliability assessment and confirmatory factor analysis to scale development procedures (Churchill, 1979; Gerbing and Anderson, 1988). Arguably, CTT is *the* dominant paradigm for addressing measurement problems in marketing research.

However, the literature on measurement and psychometrics has grown increasingly diverse. New measurement approaches have been developed. New methods, parametric and nonparametric, have been proposed. Even by the mid-1980s, Lewis (1986) had concluded that a researcher could choose from over 50 different measurement models depending on the researcher's particular needs and inclination (Thissen and Steinberg, 1986). There is little evidence to suggest a slowdown in the proliferation of measurement approaches. Unfortunately, although these developments have swept the areas of educational psychology and psychometrics, the marketing literature has remained largely insula- ted and tended to focus exclusively on CTT-based approaches.

This paper aims to draw attention to one such alternative approach—the item response theory (IRT) approach—that appears particularly promising for addressing the abovementioned contemporary measurement problems. Although IRT is not a new methodology (early IRT work dates back to the 1900s), published applications of IRT in the marketing literature are largely conspicuous by their absence (see, however, Balasubramaniam and Kamakura, 1989; Singh et al., 1990). Perhaps this is because advances in IRT technology and its tractability for practical applications have occurred only recently, and an impression persists that IRT is theoretically complex, difficult to implement, requires huge data sets, and is incoherent in its message to applied researchers. The specific purpose of this paper is to provide an accessible review of IRT by utilizing a three-pronged strategy. *First*, I aim to elucidate IRT's principles and demonstrate its relevance by illustrating how researchers can utilize this approach to address measurement problems. However, I do not claim to provide a complete, exhaustive, and in-depth tutorial on IRT models. The vast and growing literature on these models makes such a task difficult at best. Instead, my modest aim in this paper is to stimulate the interest of marketing researchers in issues of measurement theory and to discuss a particular theory (IRT) that departs from the currently dominant paradigm in marketing (i.e., CTT). *Second*, with the aim of demonstrating IRT's potential, I maintain a stance of comparative analysis throughout the paper; IRT assumptions and principles are compared with those of CTT; IRT results are compared with those obtained from CTT for two illustrative constructs; and the different approaches for tackling the abovementioned measurement problems are highlighted. Although I include this illustration *only* to make vivid the comparison between the IRT and CTT approaches, I wish to emphasize that IRT is not a panacea for all measurement woes. Rather, IRT is just another approach that rests on different assumptions than CTT and, as a consequence, allows researchers to tackle measurement issues differently. At the same time, IRT can coexist with CTT, and in many instances, these approaches might complement each other. *Third*, I aim to show that IRT is an interesting methodology because, in the tradition of Davis (1971), it reveals insights into measurement characteristics of items, measures, and concepts that (1) CTT cannot unravel and (2) challenge conventional wisdom about methods for tackling measurement problems. I begin by drawing a distinction between measurement and substantive theory and thereafter develop a comparative analysis of CTT and IRT approaches.

2. Measurement theories vs. substantive theories

> Relationship between independent and dependent variables are the focus of. . .*substantive* [theories]. However, substantive research constitutes only one part of the research process. An equally important set of research issues involves the relationship between. . .measures and the concepts or constructs they are purported to assess.
>
> Schwab (1980, p. 4).

A measurement theory sets down rules—referred to as *correspondence* rules—for linking empirical observations (*observables*) to abstract, unobservable (*latent*) concepts[1] (Blalock, 1968; Weiss and Davison, 1981). In other words, a measurement theory describes how a measuring instrument or scale performs its measuring function—that is, how the observations obtained by using an instrument translate to a position on the latent concept purported to be measured by that instrument. For instance, in the case of role conflict (RC), a measurement theory hypothesizes how the observed scores on the eight items of Rizzo et al.'s (1970) RC scale map onto the latent continuum representing the concept of RC. Such a theory may account for the characteristics of the observed scores, of the latent continuum, and of the cognitive mechanisms that underlie an individual's response to the posited items (Lord, 1952; Lord and Novick, 1968).

In accord with Schwab's (1980) comment, measurement theories must be differentiated from substantive theories, however. The former contains hypotheses about the relationship between a latent concept and its observables, while the latter specifies hypotheses about the relationships among latent concepts. Thus, for instance, the hypothesized relationship among the concepts of RC, role ambiguity (RA), and job satisfaction is the domain of substantive theory, but understanding how the operational measures or observables relate to their corresponding concept is the concern of measurement theory.

At a more abstract level, however, measurement and substantive theories share some common elements. Both specify conceptual models that allow development of formal hypotheses for empirical investigation. Because measurement theories are not well understood in marketing research, the preceding elements (i.e., hypotheses and models) are more easily recognized in a substantive context. For instance, in the case of RC, Kahn et al. (1964) posited the *role-episode model* (also see King and King, 1990) that specifies (1) definition of the RC concept itself, (2) other concepts that are expected to relate to the RC concept as either antecedents or consequences, and (3) the nature and form of the relationships among the antecedents, RC, and its consequences (referred to as the nomological net). This role-episode model in turn serves as the foundation for developing specific *hypotheses* concerning the antecedents and consequences of the RC concept that can be empirically tested.

A parallel notion of a conceptual model and hypotheses is relevant for a measurement theory. For instance, the conceptual model in a measurement theory may involve (1) characteristics of observed scores including their number, their complexity, and their scale properties, (2) properties of latent variable(s) thought to be measured by the observed scores, including dimensionality and scale properties, and (3) a mathematical model to represent the relationship between latent variables and observed scores. Here, the *observed* scores are data elements that are obtained through some operational mechanism, including an individual's responses on a scale, a supervisor's ratings of one or more subordinates, a key informant's assessment of an organizational property, or measures secured via some observational device. In contrast, *latent* variables are unobserved organizational and/or individual properties that are often inferred on the basis of some observed variables. Finally, the mathematical model is usually consistent with some assumptions about the response process that individuals or informants utilize in developing a response to the posited questions or statements. Table 1 lists various characteristics of a measurement theory (see first column) and specific properties for the RC and RA concepts implied under different measurement theories (i.e., IRT and CTT). I use the RC and RA concepts to illustrate the principles of a measurement theory and highlight the characteristics along which different measurement theories differ. However, the principles are general and can be applied broadly. Next, I discuss IRT and CTT measurement theories in the context of RC and RA concepts.

3. IRT vs. CTT: underlying conceptual model and hypotheses

Before discussing the characteristics and differences summarized in Table 1, it may be useful to provide a brief discussion of the illustrative concepts and their suitability for comparative analysis. Although any of the marketing concepts could have been utilized to compare IRT and CTT principles, several reasons favored the choice of RC and RA concepts (Kahn et al., 1964; Rizzo et al., 1970). First, both constructs are based on a rich and long tradition of theoretical and empirical work (e.g. Kahn et al., 1964; Biddle, 1986; Jackson and Schuler, 1985; Ford et al., 1975; Whetten, 1978; Pearce, 1981; King and King, 1990). In addition, these role constructs have attracted extensive empirical research, including two metaanalysis (Fisher and Gitelson, 1983; Jackson and Schuler, 1985) and several critical reviews (Schuler, 1977; Pearce, 1981; Van Sell et al., 1981; King and King, 1990). Second, RC and RA have continued to retain their importance in the marketing researcher's arsenal of theoretical constructs that can be used to study the impact of organizations on individuals and role occupants' behavioral and psychological responses (Singh, 1993; Churchill et al., 1985; Fry et al., 1986; Michaels et al., 1987; Behrman and Perreault, 1984; Edwards, 1992). Third, measurement of these constructs has elicited persistent and, in some instances, trenchant criticism (Schuler, 1977; Tracy and Johnson, 1981; House

[1] Following Kerlinger (1986, Chap. 2), I distinguish between the terms "concept" and "construct." Specifically, "construct" is utilized to refer to a "concept" with added meaning, in that it indicates that a deliberate and conscious attempt has been made to define, specify, and operationalize the focal abstract phenomenon (i.e., the "concept") for the purpose of scientific study.

Table 1
Measurement theories: illustrations of differences in models and characteristics often implied under factor analytic CTT and 2PL IRT models for the RC and RA concepts[a]

Elements of measurement theory	CTT		IRT	
	RC[b]	RA[b]	RC[b]	RA[b]
Characteristics of observed variables				
Number	Eight items	Seven items	Eight items	Seven items
Complexity	Simple	Simple	Simple	Simple
Scale property	**Interval**	**Interval**	**Ordinal**	**Ordinal**
Characteristics of latent variables				
Dimensionality	Unity	Unity	Unity	Unity
Scale property	Interval	Interval	Interval	Interval
Mathematical model for relationships				
Form	**Linear**	**Linear**	**Nonlinear**	**Nonlinear**
Parameters	λ[c]	λ[c]	a and b[d]	a and b[d]
Equation[e]	$X=\lambda T+\varepsilon$	$X=\lambda T+\varepsilon$	$P(\theta)=[1+\exp[-a(\theta-b)]^{-1}$	$P(\theta)=[1+\exp[-a(\theta-b)]^{-1}$

[a] The key differences are in bold for clarity.

[b] The specific conceptualization and operational measure utilized here is based on Kahn et al. (1964) and Rizzo et al. (1970).

[c] Parameter λ refers to a factor loading in a common or confirmatory factor analysis.

[d] Parameter a is referred to as the discrimination or sensitivity parameter and is analogous to parameter λ in CTT. However, parameter b is a threshold or affectivity parameter and is introduced to account for ordinal response categories.

[e] For more details on the equations, see Appendix B.

et al., 1983; McGee et al., 1989). King and King (1990, p. 62) went as far as to conclude, "We firmly believe that partial responsibility for the inconsistencies in research findings about RC and RA is due to deficiencies in measurement." Fourth, several researchers have called for utilizing IRT methods to examine the measurement issues surrounding theses constructs (King and King, 1990; Singh and Rhoads, 1991), yet no empirical study has tackled these issues. Below, I provide a brief background on RC and RA concepts, constructs, and their measures, explain the sources of data for the current study, and highlight key measurement concerns.

Concept definitions. Both RC and RA can be defined either from an *objective* (i.e., verifiable conditions in the work environment) or *subjective* (i.e., experienced psychological evaluation) standpoint. However, the subjective assessment has received the most attention. Thus, *subjective RC* is defined as the role occupant's evaluation concerning the degree of incompatibility of expectations associated with different role senders (e.g., boss, customers, and family). In other words, RC arises when a role occupant believes that compliance with the expectations of one role sender would make it more difficult to comply with another role sender's expectations. The *subjective RA* is defined as the role occupant's evaluation of the degree to which clear information is lacking about (1) role expectations, (2) methods for fulfilling role expectations, and/or (3) the consequences of role performance.

Operational measures. The RC and RA constructs were measured in the study by the specific scales developed by Rizzo et al. (1970). These scales have been used extensively in the literature. In fact, Jackson and Schuler (1985) report that 85% of the studies that they metaanalyzed had utilized the Rizzo et al. measures. In these measures, RC is assessed by eight items and RA by seven items (see Appendix A for item descriptions). Each item has a five-point 'strongly agree–strongly disagree' Likert scale.

Measurement concerns. Two concerns dominate RC and RA research. *First*, persistent questions have been raised about the discriminant validity of Rizzo et al.'s measures. In other words, the question is this: Do the RC and RA scales measure two unique, distinguishable constructs, or are the items in both measures simply different indicators of one general construct (e.g., role stress)? Although several empirical studies have been conducted to address this question (Tracy and Johnson, 1981; Schuler et al., 1977; McGee et al., 1989), at best, the answer has been equivocal, with empirical evidence forthcoming on both sides of the debate. *Second*, the fidelity of Rizzo et al.'s measures has been repeatedly challenged. That is, researchers ask if these measures are able to capture the entire domain of the RC and RA constructs. King and King (1990, p.53) argued that Rizzo et al.'s measures 'fall short in adequately sampling the richness and comprehensiveness of (their) theoretical content domains.' To date, this issue has not been empirically resolved.

Data source. The data utilized here to illustrate IRT analysis for RC and RA constructs involved 472 responses from salespeople selected from a list of the members of the association of Sales and Marketing Executives (referred to as the SME sample). The profile of the SME sample is as follows: 72% male, 41–45 years of median age, 4–5 years of median experience on the job, and over 60% earn above US$50,000 yearly.

Next, I turn to a discussion of Table 1.

3.1. Characteristics of observed variables

The observed variables have three characteristics. The *first* characteristic is the number of indicants, items, measures, or *observables* utilized to measure the focal concept(s). For the RC and RA concepts, there are eight and seven observables, respectively, in the operational scales developed by Rizzo et al. (1970). The *second* attribute is the complexity of each observable in terms of the number of latent variable(s) it purports to measure. For instance, the complexity level of each RC/RA item is "simple," since each item is hypothesized to measure one and *only* one latent variable. Readers can contemplate instances in which an observable might be relatively complex, providing data on multiple latent constructs (e.g., as in multitrait, multimethod designs). Finally, the *third* feature is the scale property. Following Stevens (1946), I explicitly recognize that an observable can have nominal, ordinal, interval, or ratio scale characteristics (Gaito, 1980; Michell, 1986). In a nominal scale, the numbers merely identify a specific object (e.g., social security number); an ordinal scale ranks different objects (e.g., in gradations of more or less); and an interval scale ranks objects in such a manner that a difference between ranks is constant. Lastly, a ratio scale possesses a natural or absolute zero point in addition to interval scale properties.

In comparing IRT and CTT characteristics for observed variables, Table 1 reveals that the scale property is a source of distinction. Recall that each RC and RA item was measured by using a five-point "strongly disagree–strongly agree" Likert-type response scale (coded from 1 to 5). In using CTT, researchers assume that such response scales involve interval-type data so that the computation of means, correlations, and other inferential statistics (e.g., reliability coefficients) is justified (Lord and Novick, 1968). In practice, this assumption is rarely tested before applying CTT methods. In contrast, IRT does *not* require that researchers presuppose that Likert-type data possess interval properties. Rather, in using IRT, response categories corresponding to such phrases as (1) strongly disagree, (2) disagree, (3) neither agree nor disagree, (4) agree, and (5) strongly agree or to variations of this general theme can be assumed to involve *categorical, rank-ordered* data. As discussed later, this distinction has important implications for the mathematical model underlying CTT and IRT.

3.2. Characteristics of latent variables

The latent variables are characterized by two attributes, namely dimensionality and scale property. *Dimensionality* refers to the number of distinct theoretical "factors" that are hypothesized to underlie a set of observables. This attribute follows from the principle of local independence according to which variances and covariances among a set of observables can be attributed to *only* three sources: (1) systematic variance on account of one or more (usually prespecified) latent variables, (2) unique variance specific to each observable, and (3) random error variance (McDonald, 1982; Lord and Novick, 1968). The specification of the number of latent variables under the first condition above constitutes a hypothesis for the dimensionality attribute. For RC and RA, each role concept is conceptualized as unidimensional, so that a single latent "factor" is hypothesized to underlie the observed responses. Note that this hypothesis follows from a *substantive* theory about the individual role concepts; as such, this hypothesis remains invariant across different *measurement* theories.

The *scale property* of latent variables is akin to the scale characteristic of observables. Here, the hypothesis involves whether latent variable(s) have nominal, ordinal, interval, or ratio properties. In general, this hypothesis is likely to stem from the substantive theory about the nature of the latent variables. For RC and RA, the role theory espoused and refined by Kahn et al. offers an insight into these role concepts. Both RC and RA represent a role occupant's perceptions about the pressures and expectations of role senders; however, as perceptual concepts, roles may be indexed by the degree to which they contain more or less RC and/or RA. In this sense, the latent variables of RC and RA are continuous random variables with interval-type scale properties. As before, this substantive hypothesis remains invariant across measurement theories.

Taken together, the hypotheses for observed and latent variables help delineate areas of similarity and points of distinction between CTT and IRT approaches. While the CTT draws correspondence rules between intervally scaled observables and intervally scaled latent variables, the IRT maps ordinally scaled observables onto intervally scaled latent variables.[2] This mapping or correspondence is captured by a mathematical model.

3.3. Mathematical model

Three characteristics define a mathematical model that specifies the relationship between observables and a latent variable. First, the form of the model is considered. Notably, although CTT hypothesizes a linear function, IRT posits a nonlinear function. Readers will note that a nonlinear function is more general and subsumes a linear relationship. Under IRT, the nonlinear function between the responses to an individual item and the underlying latent variable is referred to as the item response function (IRF). A variety of IRT models exist for different types of observed variables, contexts, and assumptions about the response process (Thissen and Steinberg, 1986). For instance, in the case of dichotomous response category data, potential IRT models include the Rasch or one-parameter logistic model (Rasch, 1960; Wright and Stone, 1977; Hambleton and van der Linden, 1982), the two- and three-parameter logistic model (Birnbaum, 1968; Lord and Novick, 1968), and nonparametric model (Mokken and Lewis, 1982). Corres-

[2] Some researchers have argued that the level of measurement associated with the monotonic IRT models discussed here (see Section 3.3) is no more than ordinal (van der Linden and Hambleton, 1997).

ponding models are available for polytomous response category data as well (Bock, 1972; Masters, 1982; Samejima, 1969). In addition, "unfolding" IRT models are available that do not assume monotonically increasing IRFs; instead, such models involve single-peaked, nonmonotonic IRFs (Roberts et al., 2000; Van Schuur and Kiers, 1994). The choice of a specific model is often guided by multiple concerns including the scale property of observed variables (e.g., either nominal or ordinal), the purpose of the study at hand (e.g., analyzing attitudinal or intelligence testing items), and the data demands of individual IRT models (see van der Linden and Hambleton, 1997; McDonald, 1999). To make this analysis relevant, I focus on an IRT model that would be appropriate for analyzing responses to attitudinal or personality-type scale items. Such scales have dominated marketing research to date, and their position is likely to remain unchallenged into the near future (Bearden and Netemeyer, 1999). To keep the discussion simple and focused on illustrating IRT principles, I selected an IRT model that is appropriate for observed responses obtained on a dichotomous agree/disagree scale but that can be extended to multicategory, polytomous, Likert-type response scales. Compared to dichotomous response data, IRT models for polytomous response data pose significantly greater demands in terms of large samples, limited item pools, or both (Drasgow and Hulin, 1990). However, the future availability of efficient estimation methods is likely to reduce the incremental data demands for polytomous response data. For the preceding conditions, an appropriate IRT model is the two-parameter logistic (2PL) model for dichotomous response data (Birnbaum, 1968; Reiser, 1981), which has been extended to polytomous response data as Samejima's (1969) graded response model.

Second, the mathematical model is characterized by its parameters. For a typical CTT model, such as the common factor model, a single parameter, referred to as a factor loading (λ), is specified to capture the association between the observables and the underlying latent variable. In contrast, the 2PL IRT model specifies two distinct parameters to model the observables–latent variable relationship. These parameters are referred to as the discrimination or sensitivity parameter (a_i) and the threshold or affectivity parameter (b_i). Although I interpret the IRT parameters at a later point in the paper, it is noteworthy that the a_i parameter is analogous to the λ parameter in CTT (Hulin, Drasgow, & Parsons, 1983). As such, the IRT model hypothesizes an additional parameter (b_i) for which there is no parallel in CTT. This additional parameter often yields additional information about the relationship between individual observed variables and the underlying latent variable.

Third and finally, the mathematical model is described by a specific equation utilized to operationalize the correspondence rules embodied in a measurement theory. In CTT, a typical model is the common factor model. For this model, the observed score (X) is related to the true score (T) via a factor loading (λ) as follows:

$$X_i = \lambda_i T_i + \varepsilon_i \tag{1}$$

where i indexes an item or observable and ε represents the random error component. In other words, the regression of X on T is linear, with a slope of λ (see Appendix B for more details). In contrast, the 2PL IRT model defines a nonlinear function that relates the probability of agreeing with a specific item [$P_i(\theta)$] to the underlying latent variable (θ) and via item parameters a_i and b_i. Although early IRT models utilized the normal ogive function, currently, a logistic function is commonly employed to model the relationship between the probability of checking the "agree" or "yes" category and the underlying attitude as follows.

$$P_i(\theta) = \frac{1}{1 + \exp[-a_i(\theta - b_i)]} \tag{2}$$

It may be noted that the observed X_i does not figure directly in this definition, only the probability of agreeing does (see Appendix B for more details). In this sense, IRT models are probabilistic models in which it is assumed that an individual's response to a specific item is inherently a stochastic process. Also noteworthy is that IRT models include an expression for the difference between the individual respondent's score on the latent construct (θ) and item characteristic (b_i) on a common scale. This allows for joint scaling of stimulus (items) and respondents on a common metric. Although it is evident that CTT and IRT are different measurement theories owing to their underlying hypotheses and mathematical models, the implications of these differences for key measurement issues are less clear. To clarify these implications and provide concrete comparative analysis, I next analyze the RC and RA scales by CTT and IRT approaches.

4. IRT vs. CTT: comparative results from RC and RA constructs

> We must also admit the possibility that our measurement theories. . .may not be equally valid in all settings. Perhaps simpler models will not be misleading in some settings but would produce serious biases in others. As a general rule, the more diverse the settings and more indirect the measurement, the more complex our measurement theories must be.
>
> Blalock (1984, p. 57).

Initially, I examine the hypothesis common to CTT and IRT models for the RC and RA scales. To be specific, both models hypothesize unidimensionality (Table 1). In addition, the IRT model used here assumes that the RC and RA scales are conditionally independent, which implies unidimensionality (see Appendix B). I recognize that, in some contexts and for some item content, unidimensionality is not sufficient for conditional independence, as other dependence structures may be exhibited (Bradlow et al., 1999). Although different approaches are available for testing the consequences of unidimensionality (Zhang and Stout, 1999), I sought converging evidence by using two approaches that are commonly utilized in CTT and IRT tradition: (a) parallel analysis (Horn, 1965) and (b) subset method (Bejar, 1980). A parallel analysis involves (1) obtaining "actual" eigenvalues from the correlation matrix of scale items, (2) computing "parallel" eigenvalues from a correlation matrix of equal number of random items based on equal size of sample as the actual data, and (3) comparing the actual and parallel eigenvalues to determine the number of actual eigenvalues that exceed the highest parallel eigenvalue. This comparison yields evidence concerning the dimensionality of the scale items (Humphreys and Montanelli, 1975). Typically, multiple sets of random data are generated, and expected values of parallel eigenvalues are obtained to avoid any idiosyncratic effects. By contrast, the subset method involves (1) splitting the total item set into an arbitrary number of subsets such that each subset contains a relatively homogenous pool of items, (2) estimating the threshold parameters for each item separately on the basis of the total set and its subset, and (3) plotting the pairs of threshold estimates obtained (i.e., total set vs. subset). Bejar (1980) suggested that the unidimensionality of an item set is supported if the plot of threshold parameters follows a straight line with a slope of unity. This slope is equivalent to the correlation between the two sets of IRT parameters. Below, I implement both procedures, starting with the parallel analysis.

The four highest eigenvalues for the correlation matrix of the combined set of 15 RC and RA items are 5.98, 1.95, 1.04, and 0.91. By comparison, the analytically derived parallel eigenvalues are 1.36, 1.27, 1.21, and 1.15. Because only two actual eigenvalues exceed the highest parallel eigenvalue, this suggests that the combined set of RC and RA items measures two distinct factors. Nevertheless, the first actual eigenvalue of 5.98 indicates the presence of a dominant higher-order factor that captures 40% of the variance in the combined set of RC and RA items (Hattie, 1985). To confirm the evidence of unidimensionality, I separately analyzed the RC and RA items as well. For the eight RC items, the first four actual eigenvalues are 3.91, 0.99, 0.81, and 0.57, while the corresponding parallel eigenvalues are 1.51, 1.06, 0.99, and 0.93. This comparison clearly supports the unidimensionality of RC items. Likewise, in support of the unidimensionality of the seven RA items, the four largest eigenvalues are 3.96, 0.84, 0.78, and 0.52, with corresponding parallel values of 1.11, 1.02, 0.95, and 0.89. Significantly, the first eigenvalue explains over 48% of the total variance in RC items and over 56% of the total variance in the RA items. This pattern of eigenvalues provides further support for the unidimensionality of the RC and RA items.

In order to implement the subset method, I split the eight RC items into odd (Items 1, 3, 5, and 7) and even subsets (Items 2, 4, 6, and 8; see Appendix A). Using the computer program MULTILOG, I fitted a 2PL IRT model to the items in each of the subsets and to the total set of RC items. In accord with Bejar (1980), the corresponding pairs of threshold parameters estimated for each RC item from the subset vs. the total set were plotted as shown in Fig. 1. I followed the same procedure for RA items where the odd subset contained four items (Items 1, 3, 5, and 7) while the even subset included the remaining three items. The plot of subset vs. total set threshold parameter estimates is also displayed in Fig. 1. To supplement a visual inspection of Fig. 1, I also computed a Pearson correlation between the corresponding pairs of RC and RA threshold parameter estimates. Finally, because the parallel analysis indicated the presence of a single, dominant higher-order factor in the combined set of RC and RA items that explains over 40% of the total variance, I conducted an additional subset analysis with the total set of RC and RA items and two subsets composed of either the RC items or the RA items. Drasgow and Parsons (1983) noted that IRT parameter estimates are robust to multidimensionality when a dominant, single higher-order factor underlies a set of multidimensional items. This latter analysis allowed examination of the robustness of IRT estimates for the combined set of RC and RA items.

Fig. 1 provides evidence in support of both the unidimensionality of RC and RA measures and the robustness of IRT parameter estimates obtained from the combined set of RC and RA items. The Pearson correlations for the corresponding pairs of threshold parameter estimates from the total set and subset are .97 and .99 for the RC and RA measures, respectively. In accord with this finding, the graphs in Fig. 1a and b indicate that the corresponding pairs fall closely along the 45° line, representing perfect correspondence. Moreover, the correlation for the corresponding pairs from the combined set of RC and RA items and the individual subsets is .99. Consistent with this, Fig. 1c provides evidence for the robustness of IRT estimates obtained from the combined set. As noted earlier, this coheres with the findings of Drasgow and Parsons (1983). Taken together, the preceding findings suggest that the measures do not violate underlying assumptions of CTT and IRT analysis and may be appropriate for a comparative study. I begin with results from a CTT analysis.

4.1. Analysis of RC and RA scales using the CTT approach

Because the factor structure underlying the RC and RA measures is well specified, confirmatory factor analysis was used to implement the CTT approach. I estimated a measurement model of two underlying factors with the eight RC items allowed to load freely on one factor and the seven RA items allowed to load freely on the second factor. The relationship between the underlying factor and the RC/RA measure is modeled as per the CTT equation in Eq. (1) and displayed in Table 1. No cross-loadings were modeled, and the two factors were allowed to correlate freely. The

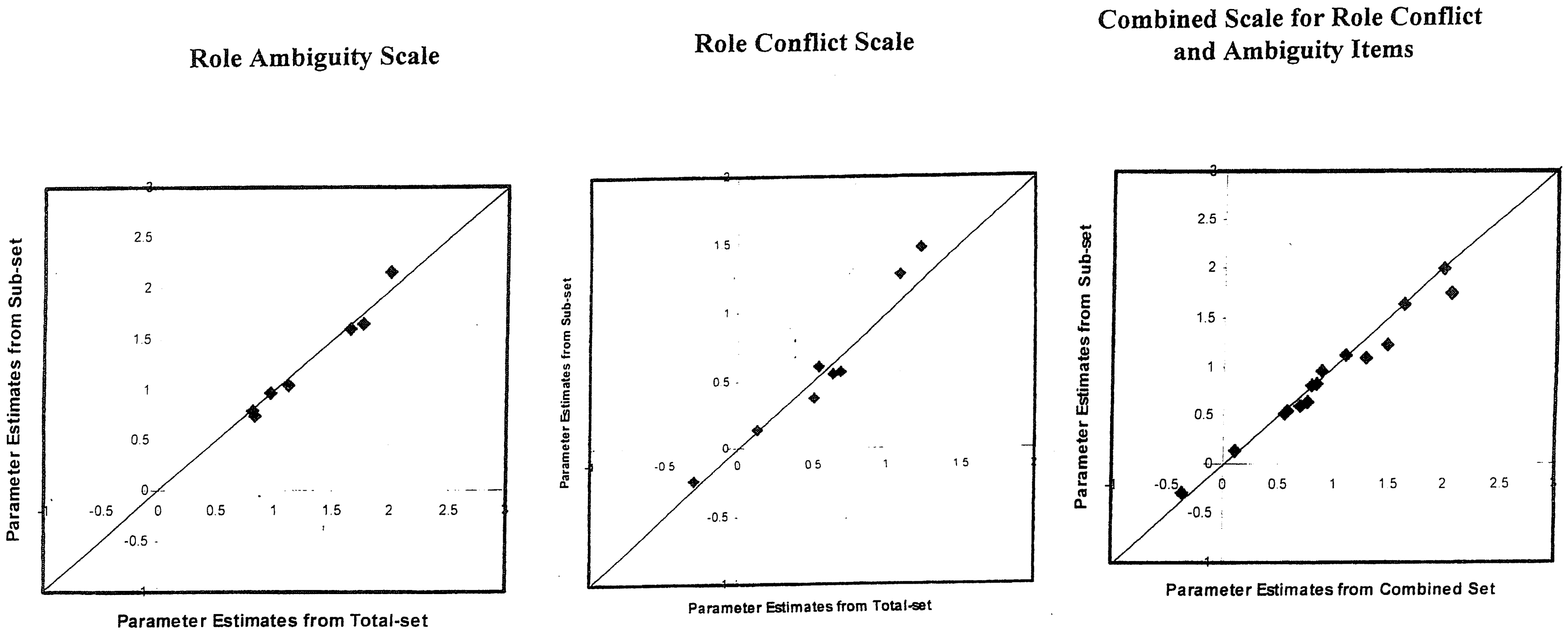


Fig. 1. IRT assumptions: testing for evidence of unidimensionality of Rizzo et al.'s RC and RA constructs.

software (EQS) was used to obtain CTT estimates via the elliptically reweighted least squares (ERLS) procedure (Bentler, 1995). The results are in Table 2.

Table 2 reveals that the CTT model fits the data for RC and RA measures reasonably well. Although the χ^2 statistics is significant ($\chi^2 = 280.4$, $df = 89$, $P < .01$), indicating that the reproduced covariances differ statistically from the observed covariances among the combined set of RC and RA items, this statistics is known to be biased for sample sizes exceeding 200 (here, $N = 472$). Consequently, I focus on other indicators to judge the goodness-of-fit. The relative fit indicators, such as the comparative fit index (CFI) and normed fit index (NFI), evaluate the degree to which a hypothesized model is an improvement over a null model, with values exceeding 0.95, indicating reasonably well-fitting models. The absolute fit indicators, such as the standardized root mean square residual (RMSR) and root mean square error of approximation (RMSEA), provide an indication of discrepancy between the observed and reproduced covariances, with values from less than 0.05 to 0.08, suggesting well-fitting models (Marsh et al., 1996). Finally, the non-normed fit index (NNFI), also known as the Tucker–Lewis index, is a measure of model parsimony balancing incremental fit with the *df* such that values exceeding 0.95 indicate parsimonious models (Marsh et al., 1996). Although the different fit indicators emphasize different aspects of model fit, Table 2 reveals that the hypothesized two-factor model meets or exceeds the goodness-of-fit criteria for each of the indicators discussed. These observations provide confidence in the CTT model fitted to the RC and RA data.

Table 2
CTT results: analyzing RC and RA scales by utilizing the confirmatory factor analysis procedures[a]

Item	Factor 1[b]	Factor 2[b]
RC1	0.62 (0.050)	
RC2	0.78 (0.048)	
RC3	0.70 (0.050)	
RC4	0.71 (0.050)	
RC5	0.58 (0.052)	
RC6	0.79 (0.047)	
RC7	0.44 (0.054)	
RC8	0.50 (0.053)	
RA1		0.46 (0.053)
RA2		0.67 (0.049)
RA3		0.38 (0.054)
RA4		0.80 (0.046)
RA5		0.89 (0.049)
RA6		0.85 (0.045)
RA7		0.78 (0.047)
Interfactor correlations		
Factor 1	1.00	
Factor 2	.58	1.00
Goodness-of-fit statistics		
χ^2 (*df*)		280.4 (89)
Comparative fit index		0.96
Normed fit index		0.95
Non-normed fit index		0.96
Standardized root mean square residual		0.05
Root mean square error of approximation		0.07
90% confidence interval		0.062–0.081
Cronbach's α reliability estimate	0.85	0.86

[a] The confirmatory factor analysis was implemented using the ERLS estimation procedure in EQS.

[b] Estimated factor loading coefficient with standard error in parentheses. All coefficients are significant at $P = .05$.

Moreover, Table 2 yields additional evidence in support of the CTT model. Each RC/RA measure has a theoretically meaningful, statistically significant, and substantial loading on its corresponding factor (all loadings > 0.35, $P < .01$). Recall that the cross-loadings have been constrained to be zero in accord with the underlying theory of RC and RA measures. This pattern of loadings supports the convergent validity of the RC and RA items. The RC and RA factors are estimated to correlate at .58, indicating that the two factors possess discriminant validity. Finally, each of the role factors has an estimated Cronbach's α reliability that exceeds .70, indicating that the measures capture significant systematic variance (Nunnally, 1978).

4.2. Analysis of RC and RA scales using the IRT approach

A 2PL IRT model was estimated for eight RC and seven RA items using MULTILOG software (Thissen, 1991). Each of the RC and RA responses was dichotomized, with respondents who agreed with an item or responded at the scale's midpoint coded as 1 and respondents disagreeing with the item coded as 0. Post hoc dichotomization of a multiple-category response scale can result in loss of information and bias (e.g., consistency with actual responses on a dichotomous response scale; also see Cohen, 1983), but I chose this approach to keep the discussion simple and focused on IRT principles. As noted earlier, the 2PL IRT model has a counterpart in the graded response model for polytomous response data, and the principles discussed here can be easily extended. MULTILOG utilizes a marginal maximum likelihood (MML) method for estimating model parameters that appears to work well with the small to moderate-sized samples common to much marketing research ($N = 100-500$). Appendix B provides some details on estimation issues for IRT models. Advanced discussions are available in McDonald (1999), Fischer (1995), and van der Linden and Hambleton (1997). The specific program lines utilized to run the MULTILOG software are provided in Appendix C.

The estimated a_i and b_i parameters for the 2PL IRT model and overall fit statistics are given in Table 3. Overall, the IRT model yields a G^2 statistics of 1738.2 with $df = 32{,}737$. This goodness-of-fit statistics is a likelihood ratio χ^2 statistics based on the ratio of observed and expected frequencies (Thissen, 1991). Although this G^2 statistics is not appropriate for evaluating overall goodness-of-fit, it can be effectively used to compare different models by computing a difference statistics and evaluating it relative to the difference in the *df*. Included in Table 3 are

Table 3
IRT results: estimated MML parameters for the RC and RA items[a]

Item	a_i[b]	b_i[b]
RC1	1.19 (0.20)	0.58 (0.14)
RC2	1.59 (0.24)	0.70 (0.11)
RC3	1.16 (0.23)	1.49 (0.24)
RC4	1.17 (0.22)	1.29 (0.22)
RC5	0.97 (0.17)	0.56 (0.17)
RC6	1.75 (0.27)	0.76 (0.11)
RC7	0.87 (0.16)	– 0.37 (0.18)
RC8	0.93 (0.16)	0.11 (0.16)
RA1	0.80 (0.19)	2.07 (0.46)
RA2	1.70 (0.25)	0.85 (0.13)
RA3	0.84 (0.18)	2.00 (0.43)
RA4	3.21 (0.70)	1.64 (0.13)
RA5	3.25 (0.50)	1.11 (0.09)
RA6	3.23 (0.52)	0.89 (0.08)
RA7	2.43 (0.33)	0.80 (0.10)
Goodness-of-fit		
G^2[c]	1738.2	
df	32737	

[a] The IRT model was estimated using the software MULTILOG. For sake of simplicity, a two-parameter model was estimated in accord with Eq. (2) in the text.

[b] The standard error of the corresponding estimate is in parentheses. The a_i parameter refers to the discrimination or sensitivity parameter, while b_i is the threshold or affectivity parameter.

[c] This is a goodness-of-fit statistics and is computed as a likelihood ratio of the expected and observed frequencies. The G^2 statistics follows a χ^2 distribution.

the estimated standard errors for each parameter estimate. A test of significance (e.g., H_0: $a_i = 0$) can be constructed by dividing the parameter estimate by its standard error and evaluating the resulting number as a t statistics with ($N -$ free parameters estimated) *df*. Based on the estimated a_i and b_i parameters, the functional relationship between the probability of agreeing (y-axis) and the underlying latent variable (x-axis) can be estimated for each RC and RA item (following Eq. (2)). Fig. 2 displays the IRF for selected items. Note that IRFs are monotonically increasing S-shaped curves that are bounded between 0 and 1 along the y-axis but are unbounded along the x-axis.

The IRF provides an interpretation of IRT item parameters. Note that in Fig. 2 that the x-axis is the latent variable (referred to as θ) and any point on the y-axis is interpreted as the probability of agreeing among individuals with the corresponding level of the latent trait on the x-axis [written as $P_i(\theta)$]. As such, $P_i(\theta)$ is usually *not* considered as the response for any one given individual. The discrimination or sensitivity parameter (a_i) is proportional to the slope of the IRF at its inflection point. The greater the a_i parameter estimate, the steeper the slope and, consequently, the more sensitive (or discriminating) the item to variability along the x-axis (i.e., latent variable) in the neighborhood of the inflection point. In this sense, the interpretation of the a_i parameter parallels that of λ in CTT. An important point of distinction is that although λ is constant throughout the range of a latent continuum, a_i is defined only at the point of inflection. An examination of Table 3 reveals that a_i parameter estimates range from 0.80 (RA1) to 3.25 (RA5) and, without exception, achieve significance (all t values >2, $P < .01$). In general, RA items have larger a_i estimates than RC items, suggesting that the former are more sensitive (or discriminating) than the latter.

The threshold or affectivity parameter (b_i) is the value of θ at the inflection point of the IRF. Technically, b_i defines that value of θ at which $P_i(\theta) = 0.50$. The notion of threshold or affectivity captures the idea that some items have a higher threshold for agreement, making them more difficult to agree with than other items. For instance, Andrich (1978, p. 565) observed that it is possible to find instances where "an agree response to an item of moderate affective value is equivalent to a neutral response to an item of high affective value." Consistent with this, the greater the b_i parameter estimate, the higher the θ value at the point of inflection, the greater the item affectivity and, for a random sample, the smaller the probability of agreeing with the item for any value of the latent variable (i.e., θ). Table 3 reveals that b_i estimates range from -0.37 (RC7) to 2.07 (RA1). Because θ is a standard variable, this suggests that the affectivity of the RC and RA items lies between 0.4 standard deviations below the mean (i.e., -0.37) and two standard deviations above the mean (i.e., 2.07). As such, the preceding range of affectivity delineates the "effective" range of a set of items.[3] Specifically, the effective range of RC and RA items appears to be between -0.4σ and 2σ. More specifically, the effective range for RC items appears to be between -0.4σ and 1.5σ, while the RA items appear to be effective between 0.8σ and 2.1σ.

4.3. A comparison and assessment of CTT and IRT results

In order to facilitate a comparison between the two approaches, in Fig. 3, I overlay the response functions from CTT and IRT analysis for selected items. In so doing, I recognize that the expected value of the latent variable in CTT (T) is not equivalent to the expected value obtained in IRT (θ); however, Fig. 3 is useful for illustration purposes. For CTT analysis, the x-axis represents the underlying latent variable in standardized units and the y-axis represents the observed scores on any given item. The CTT functional relationship is as per the equation $\hat{X} = \lambda T$, such that the slope of the regression line equals the corresponding factor loading from Table 2. For instance, the functional relationship

[3] The notion of effective range is based on the following logical arguments: (1) a set of items is effective if it helps to discriminate among people with different amounts of a latent variable, (2) the ability to discriminate is parameterized as the discrimination or sensitivity parameter (a_i) in IRT, (3) the parameter a_i is not constant throughout the range of θ, (4) instead, a_i is defined at the point of inflection where $\theta = b_i$, and as one moves away from this point the slope decreases and the item sensitivity declines, (5) consequently, b_i defines the neighborhood around which a particular item is effective, and (6) finally, for a set of items, the range of b_i estimates sets the effective range.

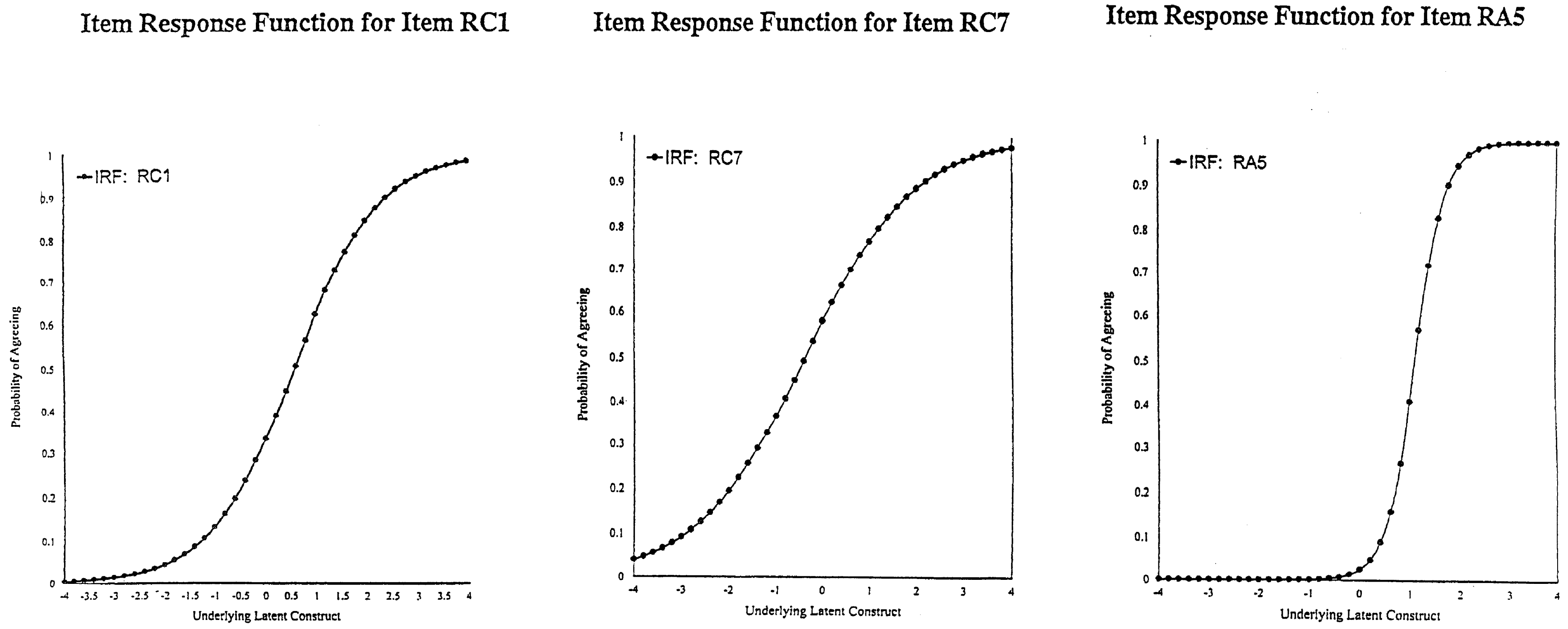


Fig. 2. IRT principles: IRFs estimated for the relationship between observed responses and latent continuum.

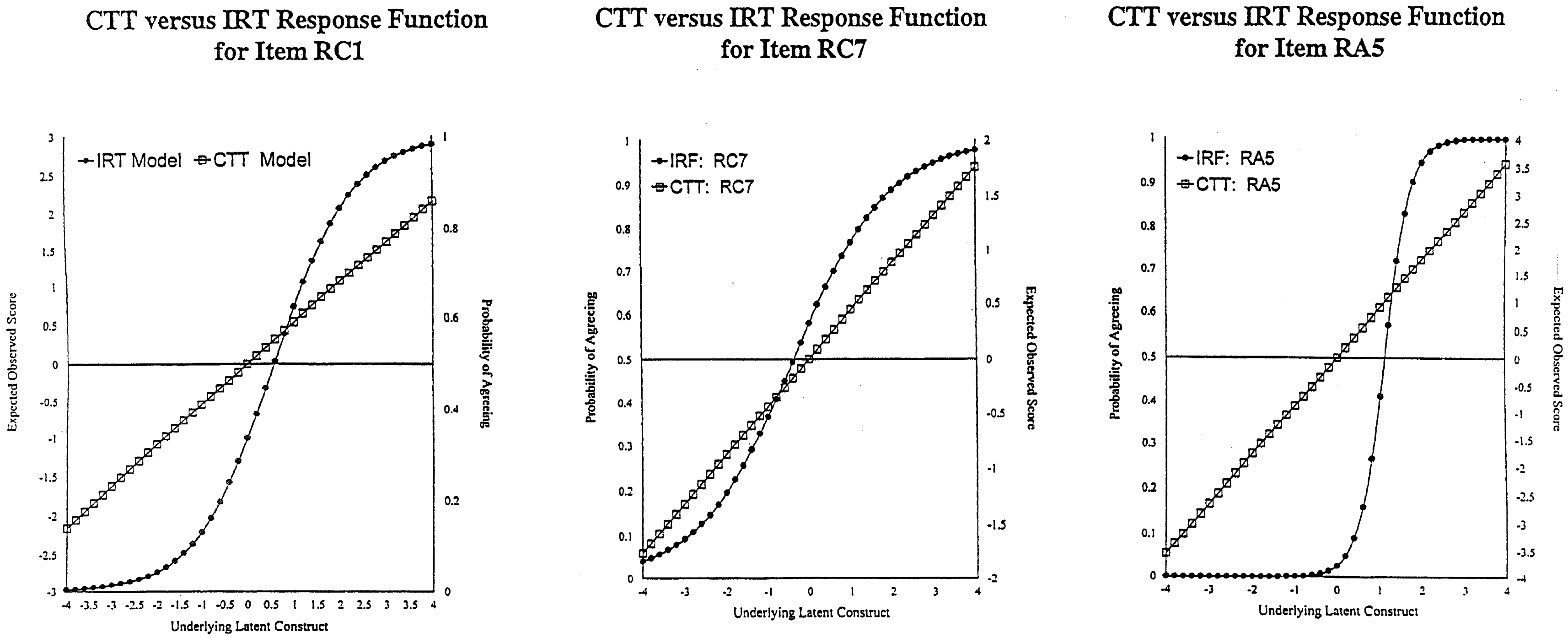


Fig. 3. CTT vs. IRT principles: response functions estimated for the relationship between expected observed responses and underlying latent continuum.

for RC1 amounts to $\hat{X} = 0.62T$. Thus, the higher the factor loading, the steeper the slope, and, consequently, the stronger the relationship between observed and true scores. At the same time, it is evident from Fig. 3 that the CTT model does not allow recognition of the bounded and categorical nature of observed responses. That is, the linear relationship implies that (1) infinite response categories are available (since the regression lines in Fig. 3 extend to positive and negative infinity on both axes) and (2) response categories are continuous numbers that allow individuals to check any incremental number along the scale for recording observed responses (for instance, say 2.3, since the *x*-axis maps onto all real values on the *y*-axis).

Three comments are noteworthy in comparing the CTT and IRT results displayed in Fig. 3. *First*, the IRT model appears more realistic in mapping the relationship between observed and latent variables. The IRT model views the observed scores as probabilities of agreeing (or disagreeing) with an item. This probabilistic view is consistent with the cognitive process that underlies individuals' responses to questionnaire items (Reiser, 1981). That is, an individual's response is not based solely on his/her "true" standing on the latent variable but is also affected by other cognitive processes that interfere with his/her response. Such cognitive processes include distracting information and events, mood swings and processes, social desirability and other biases, and random memory processes. Thus, among a group of individuals who have the same "true" standing on a latent variable, the probability of each person's agreeing with the item is neither 0 nor 1; instead, it is a finite, nonzero number between 0 and 1. Consistent with this notion, the IRT model bounds the observed response probabilities between 0 and 1. By contrast, as noted above, the CTT model is unbounded and continuous along the observed scores (*y*-axis), assuming that the responses are obtained on an intervally scaled response format. At the same time, a CTT model allows for random errors in response processes by modeling the variability along the item response regression line and including an additive component to represent measurement error (noted as ε in Eq. (1)). However, to the extent observed scores in reality (1) are often bounded by a limited number of scale categories (e.g., a five- to nine-point Likert scale) and (2) occur only as categorical options, since responses between any two categories are not scaled (i.e., a response of 1.5 is often not allowed), it appears that the IRT's probabilistic modeling is an advantage. As such, in the words of Blalock (1984), although the CTT model is simpler than the IRT model, the former may not be as appropriate as the latter to model realistic response data.

Second, the CTT and IRT models differ in terms of their parameters. Recall that the IRT model has two parameters (a_i and b_i) while the CTT model has just one (λ). Moreover, although the item sensitivity parameter (a_i) and the factor loading parameter (λ) are analogous to each other, the IRT model estimates an additional parameter to capture item affectivity (b_i). Indeed, if the RC and RA items were equally effective, the b_i parameter would be a constant across the RC and RA items. On the contrary, if item affectivity varied significantly across items, the b_i parameter estimates would provide meaningful information about item responses. The affectivity equivalence assumption can be tested in IRT analysis by constraining the b_i parameters to equal each other for all RC and RA items. I estimated a constrained IRT model for the RC and RA items and obtained a goodness-of-fit statistics G^2 (b_i's fixed) of 1974. This value compares with a G^2 (b_i's free) of 1738 for the unconstrained IRT model (see Table 3). The difference between these G^2 statistics ($G_{\text{diff}}^2 = 1974 - 1738 = 236$) provides a χ^2 statistics for testing the validity of constraining conditions (Thissen, 1991). Thus, the affectivity equivalence hypothesis that the b_i's are constant for all RC and RA items is evaluated as $\chi_{\text{diff}}^2 = 236$ with $df = 14$ (15 b_i's constrained, one b_i estimated), implying that the hypothesis is rejected resoundingly ($P < .001$). Consistent with this, the affectivity of RC and RA items varies from -0.37 to 2.07, covering a range of about 2.5 standard deviations on the latent trait scale (Table 3). Clearly, affectivity equivalence is not a tenable proposition for the RC and RA items.

Third, since a nonlinear model subsumes a linear function, the IRT model can approximate a CTT model; however, the reverse is not true. For instance, an examination of Fig. 3 suggests that, in the case of item RC7, the IRT and CTT models are reasonably similar (except at the extremes); however, in the case of RC1 and RA5, clear differences are discernable. Evidently, the linear CTT model is unable to faithfully capture the nonlinear processes underlying items such as RA5 and RC1. Because the IRT model is more general, realistic, and appropriate, the conclusion that, at least for RC and RA scales, the CTT model is statistically inappropriate appears warranted. Moreover, since there is nothing inherently unique about the way the RC and RA concepts were developed or operationalized, it is likely that the conclusion about the appropriateness of the IRT model is relevant for other marketing constructs as well.

Few would argue with the statistical merits of IRT but it is evident that the investment in IRT entails nontrivial costs, even, to some, a heavy price. This price exacted for using IRT includes learning a mathematically complex measurement theory, understanding estimation issues and software handling, greater data demands, and deriving sound and meaningful interpretations of IRT models. *Are IRT's benefits worth its price?* I believe that it is of critical importance to address this question in the context of measurement research in marketing. Put another way, before embracing IRT, marketing researchers should critically and closely examine the cost/benefit tradeoffs involved despite the measurement theory's technical merits. Clearly, it is not possible to resolve this question within the framework of a single article such as this. Instead, I will attempt to take an initial step in this direction by outlining the benefits of IRT in the context of RC and RA concepts. I hope to delineate specific instances in which the use of IRT is likely to be beneficial. However, to fully convey its benefits, I first discuss IRT information functions (IFs) that are germane to its use.

5. Key IRT characteristic: IFs and measurement precision

IRT views the precision (or standard error) of measurement from an informational perspective. This perspective departs substantially from an approach based on CTT. Typically, the precision of measurement under CTT is estimated utilizing a reliability coefficient that represents the average precision of a given scale across all respondents in a sample.[4] Several reliability estimates are available, with selection depending upon the scale of measurement, the number of items, and the research design used; examples are the Spearman–Brown coefficient, Cronbach's α, and the test–retest reliability coefficient (Zeller and Carmines, 1980). Under IRT, the precision of measurement is based on *information* [$I(\theta)$] and the *standard error* of θ [$\sigma(\theta_\varepsilon)$; Lord, 1980; Mellenbergh, 1996). Mathematically, $\sigma(\theta_\varepsilon)$ equals the inverse square root of $I(\theta)$. Conceptually, the more information one has with which to estimate θ, the smaller the error of the estimate and higher the measurement precision. Thus, information is inversely related to measurement error and positively related to measurement precision. More specifically, $\sigma(\theta_\varepsilon)$ defines a confidence interval around the estimate of the latent variable θ. The interpretation of $\sigma(\theta_\varepsilon)$ is in accord with standard statistical methods, that is, a 95% confidence interval for the latent variable is estimated as [$\theta \pm 1.96\sigma(\theta_\varepsilon)$]. Consequently, as information about $\theta[I(\theta)]$ increases, the confidence interval around θ [$\sigma(\theta_\varepsilon)$] decreases and, as a result, measurement precision increases (Lord, 1980). Interpretationally, information or measurement precision parallels the notion of reliability.[5]

The notion of information for *any* item i [i.e., $I_i(\theta)$] is itself defined as follows:

$$I_i(\theta) = \frac{\{P_i'(\theta)\}^2}{\{P_i(\theta)Q_i(\theta)\}} \tag{3}$$

where $P_i(\theta)$ is the probability of agreeing for the ith item as per Eq. (2), $Q_i(\theta)$ is the probability of disagreeing (i.e., $[1-P_i(\theta)]$), and $P_i'(\theta)$ is the first partial derivative of $P_i(\theta)$ (i.e., $\delta[P_i(\theta)]/\delta\theta$). Because the value of $I_i(\theta)$ varies with θ, the former is often referred to as the item IF (IIF).

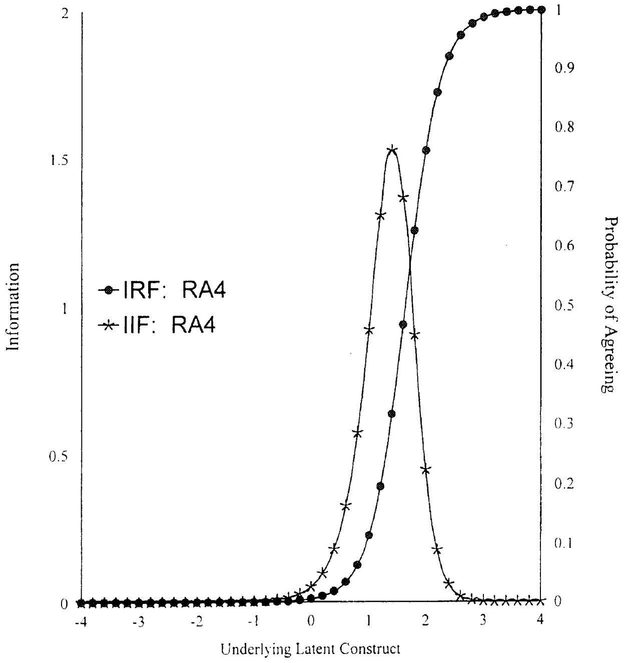


Fig. 4. IRT characteristics: IRF and IIF estimated for item RA4.

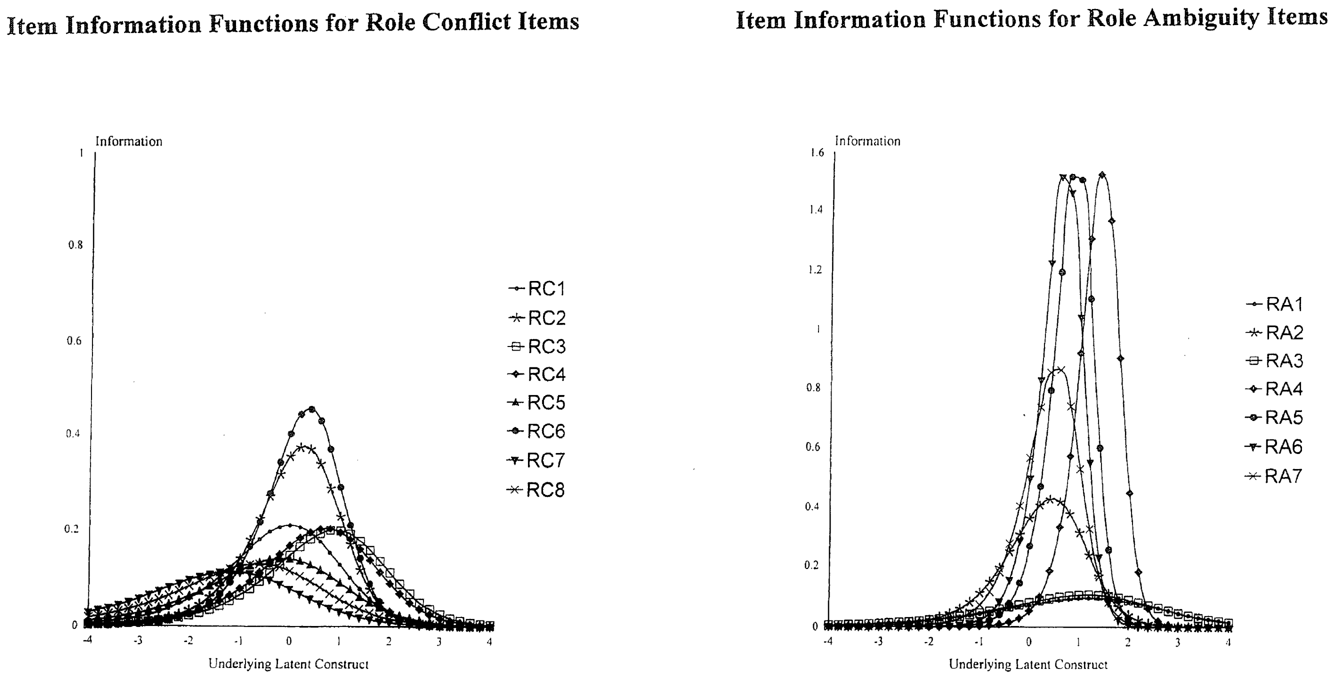
In order to illustrate the IIF of Eq. (3), Fig. 4 displays the IRF and IIF for a specific item (RA4) based on asymptotic standard error estimates obtained from MULTILOG. Fig. 5 provides IIFs separately for all RC and RA items using corresponding asymptotic SE estimates.

Fig. 4 reveals that the notion of information and measurement precision has several unique properties under IRT. First, unlike CTT reliability estimates, information for any item is *not* a constant. Rather, it is estimated as a *function* of the latent variable θ, such that different levels of θ are associated with different values of information. As an illustration, consider item RA4. Fig. 4 indicates that the information provided by RA4 is relatively *insignificant* when the latent variable is either below the midpoint (i.e., $\theta < 0.0$) or more than two and a half standard deviations above the mean (i.e., $\theta > 2.5$). However, in between this range ($0.0 < \theta < 2.5$), item RA4 provides meaningful information (values ranging from 0.1 to 1.5). In this sense, information and, consequently, measurement precision is defined locally for each and every value of θ (Lord, 1980).

Second, the shape of the IF depends solely upon item parameters. In particular, the IIF peak occurs at that point on the latent variable where θ equals the affectivity parameter (b_i) for any given item. As such, for RA4, IIF peaks when θ equals 1.64, the estimated affectivity parameter for this item (cf. Table 3). Because an IIF peak represents the most information an item can provide and, by definition, the most measurement precision it can yield, the affectivity parameter defines that point on the latent continuum around which an item is most effective. Furthermore, the height of the IIF peak is related to the sensitivity parameter (a_i). That

[4] In fact, the standard error of measurement $\sigma(\theta)$ is related to reliability ρ_{xx} by the following formula: $\sigma(\theta) = \sigma_x\sqrt{(1-\rho_{xx})}$, where σ_x is the standard deviation of observed scores (cf. Nunnally, 1978, p. 218).

[5] In fact, Green et al. (1984) have proposed a MRI based on IRT's characteristics of information and standard error of measurement. This MRI can be calculated and compared with appropriate CTT indices of reliability (e.g., Cronbach's α). Computationally, MRI is computed as $\sigma^2(\theta) - \sigma^2(\theta\varepsilon)/\sigma^2(\theta)$, where $\sigma^2(\theta)$ is the variance of the latent variable θ and $\sigma(\theta\varepsilon)$ is the standard error of measurement. Recall that $\sigma^2(\theta\varepsilon) = 1/I_i(\theta)$.


Item Information Functions for Role Conflict Items

Item Information Functions for Role Ambiguity Items

Fig. 5. IRT characteristics: IIFs estimated for Rizzo et al.'s RC and RA items.

is, the greater the estimated value of a_i, the steeper the slope of the IRF and the higher the peak value of IIF.[6] Note, for instance, that the peak value of RA2's IIF is significantly smaller than that for RA4, since the former item has an estimated sensitivity parameter ($a_{RA2}=1.70$) that is about half the estimated value for the latter ($a_{RA4}=3.21$). As such, RA4 provides significantly higher measurement precision than RA2. In this sense, the *amount* of information provided by any item is a function of its sensitivity, whereas the *location* of this information is determined by its affectivity.

Third, IRT provides IIFs for each item on the scale, independent of other items. As such, it is legitimate to expect that some items provide effective measurement on the extreme values of θ and that other items are more effective in the middle range. For instance, Fig. 5 reveals that RA1 is informationally effective for the *positive* extreme values of the latent variable (i.e., $\theta > 2$); items RA2, RA6, and RA7 are effective in the middle range (i.e., $-1 < \theta < 1$); and none of the RA items are effective for the negative extreme values of θ (i.e., $\theta < -1.5$). More importantly, the IIF characteristic of any specific RA or RC item is *not* likely to change even if additional RA or RC items are added to (or deleted from) the overall scale. This is because IIFs depend solely on item parameters that, in turn, are invariant (Dorans, 1985).

Fourth, the overall informational content of a scale of *n* items, referred to as the test or scale IF, is defined as the additive function of *n* IIFs for the individual items (Eq. (4)). Mathematically,

$$I(\theta) = \sum_{i=1}^{n} I_i(\theta) \tag{4}$$

Fig. 6 displays the IFs for the eight RC and seven RA items. Like the IIFs, the IF is inversely related to the standard error of measurement and is positively related to measurement precision; however, in this instance, these characteristics refer to the entire scale. Otherwise, the interpretation of IFs parallels that of IIFs. For instance, Fig. 6 suggests the following conclusions: (1) the RA scale has more informational content and thus yields greater measurement precision than the RC scale (i.e., peak IF_{RA}>peak IF_{RC}); (2) for $\theta < -0.10$, the RC scale is more effective than the RA scale (i.e., for all $\theta < -0.10$, $\mathrm{IF}_{RC} > \mathrm{IF}_{RA}$); (3) however, for $\theta > -0.10$, the RA scale is more effective than the RC scale (i.e., for all $\theta > -0.10$, $\mathrm{IF}_{RA} > \mathrm{IF}_{RC}$); and finally (4) both scales have poor measurement precision for $\theta > 2.50$ and $\theta < -2.50$.[7]

In comparison to the CTT approach for conceptualizing and estimating measure reliability (Nunnally, 1978), the

[6] In particular, the peak value of any item's IIF is proportional to the square of its discrimination parameter. Thus, for two hypothetical items, whose a_i parameters are such that $a_1 = 2 * a_2$, the corresponding information functions will satisfy the condition that peak $\mathrm{IIF}_1 = 4 *$ peak IIF_2.

[7] The MRI can be calculated for the RC and RA scales. For the RA scale, the peak IF value is around 4.9, so that $\sigma^2(\theta\varepsilon)$ equals 1/4.9 or 0.20. Assuming the variance of latent variable is 1 in accord with its standard units (i.e., $\sigma^2(\theta) = 1$), the MRI at the IF peak for RA is (1 − 0.20/1), which equals 0.80. However, for the RC scale, the peak IF is only 2.0. Thus, the MRI at the IF peak for the RC scale is only 0.50.

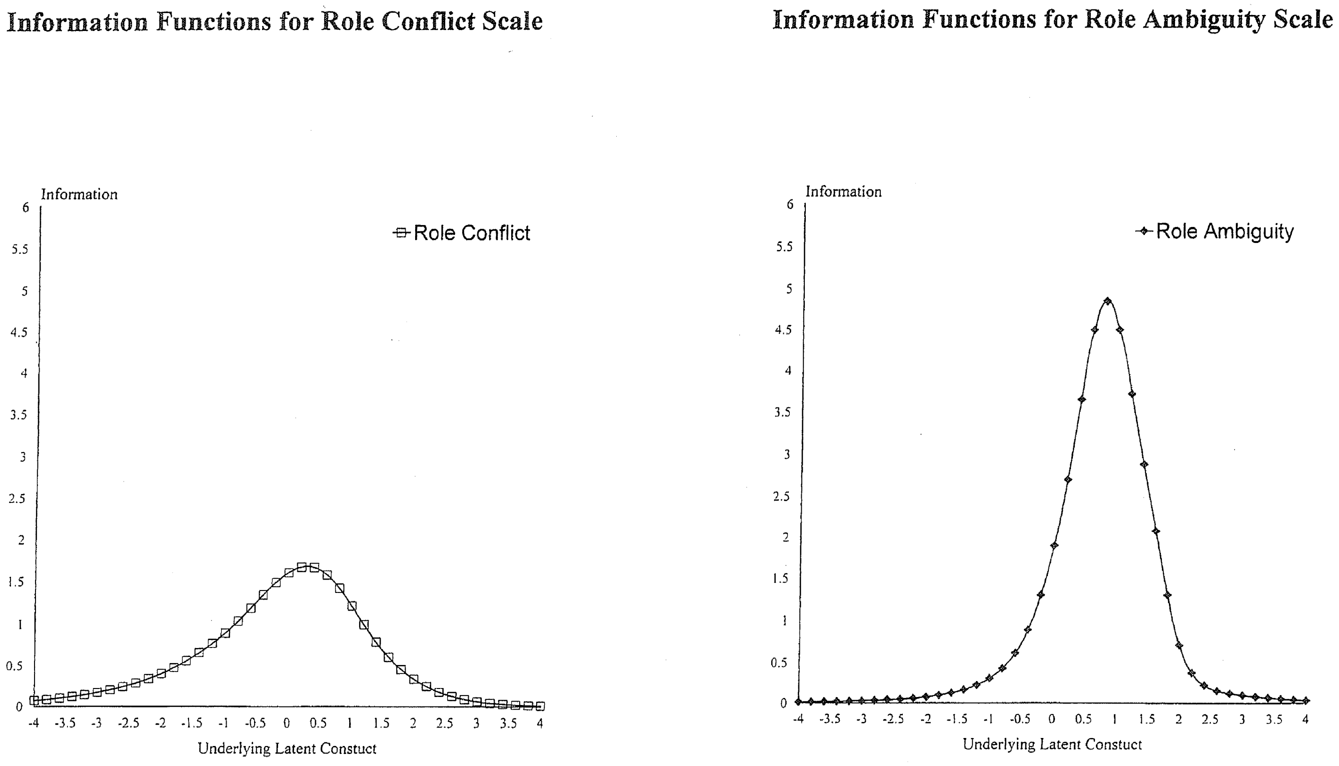


Fig. 6. IRT characteristics: IFs estimated for Rizzo et al.'s RC and RA constructs.

IRT-based IFs propose a radical departure from conventional thought. This departure is so radical that some psychometricians, like Samejima (1977, p. 243), have gone so far as to conclude that CTT-based reliability is a "dead concept in test theory." In general, comparisons between CTT and IRT approaches to measurement precision rest on four differences. First, a CTT-based reliability index such as Cronbach's α is a joint property of all of the items on the scale and the particular individuals sampled.[8] Individual items cannot be generally indexed by a reliability measure, especially in cross-sectional data. In contrast, IIFs are defined for *each* item, *independent* of other items on the scale. Second, for a given sample, Cronbach's α provides a constant estimate for measurement precision. Stated differently, the reliability of a scale of n items is unaffected by the specific level of the underlying latent variable being measured, such as the negative or positive extremes of the latent continuum. In contrast, IIFs are defined locally for each and every value of the latent variable. Third, estimates of Cronbach's α are likely to vary across samples. This is evident since α is a function of observed variance, which in turn is a function of sample homogeneity. In contrast, IIFs are theoretically invariant because they are solely dependent on item parameters. Fourth, Cronbach's α itself is valid under some very strict assumptions. Namely, that the n scale items are parallel or τ-equivalent, that is, they have equal or linearly proportional true scores (Zeller and Carmines, 1980). In addition, it is assumed that the error variances are uncorrelated across the n scale items. Although the validity of these assumptions can be empirically investigated (e.g., using LISREL), such investigation is rare in marketing research. Several researchers have observed that these requirements are difficult to meet in empirical responses (Samejima, 1977).

6. Utilizing IRT to tackle contemporary measurement problems: an illustration using RC and RA concepts

Next, I discuss how IRT can be utilized to address the four specific measurement problems noted in Section 1. I make no claim that IRT analysis invariably provides a better solution to the measurement problems than does CTT, or even that it always provides a satisfactory solution. Rather, my intention is to illustrate the comparative potential of the two approaches to solve contemporary measurement problems and enhance the tool kit of marketing researchers. Table 4 summarizes the key points discussed below.

6.1. Bandwidth–fidelity problem

In discussing the bandwidth–fidelity problems for the RC and RA constructs, King and King (1990) noted that, although the RC and RA constructs appear to have accept-

[8] For the sake of my discussion, I focus on Cronbach's α simply because this reliability measure is widely used in marketing research. However, the arguments apply to other reliability measures as well.

Table 4
IRT vs. CTT approaches: some implications for contemporary measurement issues

Measurement issue	CTT approach	IRT approach
Bandwidth–fidelity problem		
Select a unidimensional set of items that are either (a) similar to each other or (b) different from each other	Select items that maximize fidelity or reliability. No consideration given to bandwidth issues, as items are assumed to be equivalent. Often favors items that are similar to each other.	Explicitly consider bandwidth–fidelity tradeoffs. With a finite set of items, it is impossible to maximize both bandwidth and fidelity. Select a unidimensional set of items that provide information at either a defined range of the underlying construct to maximize fidelity or at different points along the construct continuum to maximize bandwidth.
Directional factors problem		
Develop a unidimensional set of items that is either (a) worded in the same direction or (b) split the set with half the items worded in the positive direction and remaining items worded in the negative direction.	Select items that are worded in the same direction to avoid factors that reflect direction of wording.	Use either option as the IRT procedures appear less sensitive to directional factors.
	Common analytical approaches (e.g., factor analysis) are sensitive to directional factors.	Include positive and negative worded items for each construct allows partialling out the effect of direction-of-wording.
Scale efficiency problem		
To reduce respondent burden and improve data quality, select a "short form" of a given scale by selecting items that either (a) preserve reliability or (b) preserve validity.	Select items with the highest factor loadings. This approach provides maximal reliability for the short form.	Short forms involve bandwidth–fidelity tradeoffs. Items may be selected for maximal fidelity by selecting items that provide maximal information in a range of interest along the underlying continuum.
	No clear guideline available for selecting items that directly preserve validity.	Select items for maximal bandwidth so that they provide maximal information at different points the construct continuum. Preceding choices will likely produce different short forms.
Scale refinement problem		
In working with a well-established construct, a researcher has to decide between (a) using the original set of items without making any alterations and/or additions even though contextual and temporal changes make the original set less relevant and (b) developing a "new" construct to tap current reflections of the phenomenon and establish its reliability and validity.	Use the original scale items without alterations. Reliability and validity indicators are properties of the entire scale. Altering or adding items can change scale properties in ways that can not be easily discerned or determined.	Use the notion of a construct item bank to add additional items to address current fidelity or bandwidth gaps.
	A "new" construct may also be developed by starting anew.	Use the information functions to identify poor or redundant items for the purposes of improving the quality of the bank by either deleting the item or altering it. Considering the purpose of the study, select specific items from the bank to provide peaked or flat scale information function.

able reliabilities in most empirical research, the operational constructs developed by Rizzo et al. (1970) are conceptually deficient as they fail to tap the underlying richness of the focal concepts. Although King and King's analysis is theoretically compelling, little empirical evidence is available from CTT analysis to suggest that the RC and RA constructs are psychometrically weak. Past empirical work coheres with my CTT results—RC and RA items load on separate factors, with strong loadings, high reliabilities, and low interfactor correlation—indicating psychometric support for the RC and RA constructs. Not surprisingly, Jackson and Schuler (1985) noted in their metaanalysis that over 85% of the past studies had used these operational measures.

In order to obtain an IRT perspective, I examined the scale and IIFs for RC and RA constructs provided in Figs. 5 and 6. This examination provides several interesting insights into *bandwidth–fidelity* issues for the RC and RA constructs. By bandwidth, I mean the range of underlying trait for which an item or scale provides reasonable measurement precision. First, the RC construct has a wider bandwidth than the RA construct since IF values for the former cover a wider range along the role stress continuum than those for the latter. Second, and in contradiction to the first, the RA construct has significantly higher *fidelity* (i.e., precision) than the RC measure since, for all $\theta > -0.10$, the IF magnitude for the former far exceeds that of the latter. Third, on the basis of the marginal reliability index (MRI), neither measure satisfies acceptable standards of reliability (i.e., >0.70). The sole exception is the RA construct when utilized in the range $\{0.25 < \theta < 1.5\}$, since the IF exceeds 3.3 (and MRI $\geq$ 0.70) *only* in this range. In sum, the RC and RA constructs suffer from serious bandwidth–fidelity problems; the Rizzo et al. constructs have small *empirical* bandwidths and lack fidelity over much of their empirical bandwidths. Thus, my illustrative IRT results provide the empirical evidence that King and King lacked in their incisive and compelling theoretical analysis.

Why did the CTT analysis fail to reveal this shortcoming? The CTT analysis is based on correlational data. When items correlate highly, they tend to produce high factor loadings and large reliability coefficients. However, correlations are *not* defined locally for different values of the underlying latent variable. Rather, they are based on aggregating across all possible values of the latent variable. Apparently, although the RC construct lacks precision at any single value of the latent continuum (i.e., MRI < 0.7, for all θ), when aggregated across all θ values, the CTT reliability appears acceptable (α>.70). Likewise, a CTT-based reliability coefficient may tell us if a measure is reliable or not, but it tells us nothing about *where* and in *what range* of the latent continuum this reliability exists. This seems to be the case for the RA construct. Clearly, the RA construct is reliable, but as the IRT analysis reveals, it provides measurement precision only in a narrow range of θ.

Is it plausible that increasing precision will invariably result in narrower bandwidth? Although the illustrative example does not directly address this issue, there is growing recognition that, while including (or selecting) similarly worded items increases reliability, such increases undermine the validity (or bandwidth) of the focal construct. This notion is consistent with Churchill and Peter's (1984, p. 370) metaanalytic results indicating that increasing reliability tended to favor selection of "items (which) were so similar (to each other) that they *underidentify* constructs." Such a finding is anomalous from the standpoint of CTT but poses no problems for IRT. Rather, IRT explains why such apparently anomalous results may arise and aids researchers by explicitly revealing the bandwidth–fidelity tradeoffs.[9]

6.2. Directional factors problem

The "direction of wording" issue poses an interesting dilemma for marketing researchers—it appears to be a good psychometric practice to include both positively and negatively worded items for each construct, but doing so often increases the dimensionality of a construct by creating unnecessary "directional" factors (i.e., corresponding to positively and negatively worded items; Richardson, 1936; Idaszak and Drasgow, 1987). Consequently, psychometric recommendations to marketing researchers have vacillated between "equal number of positive and negative items" and "all items in one direction." To date, no clear guidelines for choosing between these options exists.

The RC and RA constructs provide an interesting illustrative example in this regard. Note that all the RC items in the Appendix A are worded negatively (i.e., agreement implies greater RC), and all of the RA items are worded positively (i.e., agreement implies lower RA). This is a variation of both the recommendation to have equal numbers of positive and negative items and the recommendation to have all items in one direction. To the extent positively and negatively worded items generate directional factors, it is clear that these factors will be aligned completely along the substantive RC and RA factors. This alignment in turn is likely to lead to overestimation of the discriminant validity of the RC and RA constructs. That is, the obtained evidence for the discriminant validity of RC and RA constructs may be partly (or wholly) the result of direction-of-wording factors, depending on the sensitivity of the analytical procedures to directional factors.

The evidence of discriminant validity from CTT analysis appears unequivocal. A model that does not allow cross-loadings fits the data for RC and RA items fairly well,

[9] Readers might find the tradeoffs apparent. By selecting items that have IIF peaks around the same point on the latent continuum, it is possible to significantly enhance the fidelity of the scale, but such fidelity is *localized*. This is because the overall scale information, and hence the fidelity of the scale, is an additive sum of the individual IIFs. In contrast, if items are selected so that they peak at different points on the latent continuum, one can attain wide bandwidth, but the fidelity at any local value of θ is likely to be relatively small.

producing significant and large factor loadings. In addition, the epistemic correlation between RC and RA constructs is only .58, indicating that less than 35% of the variance is common among them. However, as noted above, the factor-loading pattern is also consistent with the positive and negative directional factors. Thus, the evidence of discriminant validity is probably inflated, and certainly confounded, by directional factors. *What if each of the RC and RA constructs had equal numbers of positive and negative items?* In this instance, it would have been possible to separately estimate the directional factors and obtain a less biased evidence for the discriminant validity of the RC and RA constructs.

Howell et al. (1988) conducted such a study. Howell et al. modified Rizzo et al.'s RC and RA constructs so that each construct had equal numbers of positively and negatively keyed items. Upon analyzing these revised measures, Howell et al. observed, "(We) provide a relatively strong indication that the (two-factor) structure that factor analysis has suggested for years with regard to these scales may be a result of method artifact rather than true differences in the RA and RC constructs as operationalized." Howell et al. went on to conclude that the RC and RA constructs as developed by Rizzo et al. lack discriminant validity, although the underlying concepts of RC and RA might well be conceptually distinct (King and King, 1990).

In this context, the IRT analysis offers interesting perspectives. Two (or more) constructs are likely to be redundant if the informational content in each scale is not unique. Such is the case with the RC and RA constructs, since a significant range along the latent continuum does *not* exist where either (1) the IF for RC is large *and* the IF for RA is small or (2) the IF for RA is large *and* the IF for RC is small. Thus, the results from IRT analysis appear to suggest that RC and RA measures lack discriminant validity. This finding coheres with my earlier analysis indicating that pooling the RC and RA items does not appear to violate Bejar's condition for unidimensionality. More importantly, to the extent that RC and RA are theoretically distinct concepts (King and King, 1990), it is apparent that Rizzo et al.'s measures fail to faithfully account for these theoretical differences.

Clearly, CTT approaches are more sensitive to directional factors than IRT procedures. In a simulation study using CTT approaches, Schmitt and Stults (1985, p. 367) reported that "regardless of data source, when only 10% of the respondents are careless..., a clearly definable negative factor is generated." In contrast, IRT analysis appears less sensitive to wording effects, as the parameter estimates remain largely invariant to pooling RC and RA items (Lord, 1980). Thus, it is probable that the evidence of discriminant validity obtained in CTT analysis is an artifact of the positive and negative wordings of items. Although more research may be necessary to resolve this issue, it is evident that, although the CTT approach falters when confronted with direction of wording artifacts, the IRT approach remains somewhat robust (see Table 4).

6.3. Scale efficiency problem

Researchers often seek "short forms" of established scales, since these enhance efficiency by reducing respondent burden and possibly enhance data quality. However, an efficiency problem arises because it is unclear if short forms compromise psychometric properties of focal constructs. This problem is evident for the Rizzo et al.'s RC and RA constructs, as several researchers have reported using short-form versions of these constructs (e.g., Singh et al., 1994). Here, I illustrate the differences between CTT and IRT approaches by considering a three-item short form for RC and RA constructs. The arguments have general applicability, however.

Under CTT, the recommended practice is to select items with the highest factor loadings, which preserves the reliability of the constructs (Bollen and Lennox, 1991). Thus, if the goal was to select a three-item efficient scale for RC and RA constructs based on the CTT results in Table 2, the choice would be items RC2, RC4, and RC6 for RC and items RA4, RA5, and RA6 for RA. These short forms produce Cronbach's α reliabilities of .80 and .88 for RC and RA, respectively; the corresponding estimates based on full scales are .85 and .86. Notably, the short form of the RA scale has higher reliability than the full scale, indicating the higher level of internal homogeneity among the selected RA items.

Under IRT, short forms can be developed using different bandwidth–fidelity criteria. For instance, if a researcher wants to preserve construct bandwidth, short-form items may be selected to provide maximal information at different points along the trait continuum. One way to accomplish this is to group the RC and RA items in accord with their affectivity, as shown in Table 5 When multiple items are available within a given range on the underlying continuum, an item with the highest sensitivity parameter can be selected to obtain maximal information. For instance, in the range of -1 to <0, there is only one RC item: RC7. However, in the range of 0 to <1, five RC items are available, of which RC1 is preferred, since it has the highest

Table 5
Grouping of IRC and RA items based on their affectivity

Range of θ[a]	Items with corresponding affectivity or b_i parameter estimates[b]
from -3 to <-2	None
from -2 to <-1	None
from -1 to <0	*RC7*
from 0 to <1	*RC1*, RC5, RC6, RC2, RC8, *RA7*, RA6
from 1 to <2	*RC3*, RC4, *RA4*, RA5
from 2 to <3	*RA1*, RA3

[a] This represents the range along the underlying construct continuum.
[b] This is based on estimates provided in Table 3.

sensitivity or a_i parameter (see Tables 3 and 5). Based on this, the choices of short-form items are RC7, RC1, and RC3 and RA7, RA4, and RA1 for the RC and RA constructs, respectively.

A comparison of RC and RA short forms obtained from CTT and IRT approaches is in Table 6. First, note that none of the three RC items selected under each approach appears on both the CTT and IRT short forms. For the RA construct, only one item (RA1) appears on both short forms. Second, the CTT approach preserves reliability, but the IRT approach does not. The IRT short forms produce reliability estimates of .60 and .69 for the RC and RA constructs, respectively; the corresponding values for the CTT short forms are .80 and .88. Consistent with this pattern, the interitem correlations for the CTT short forms are higher than the corresponding interitem correlations for the IRT short forms (see last row, Table 6). Third, both short forms correlate equivalently with their full scales. That is, both the CTT and IRT short forms for RA correlate at .93 with the full RA scale. The RC short forms do likewise. Fourth, the correlation between the CTT and IRT short forms is not high. For instance, the RC short forms correlate only at .67, and the RA short forms correlate at .84. Finally, the correlation between the RC and RA constructs is lower for the IRT short forms (.37) than for the CTT short forms (.46).

Overall, as noted in Table 4, the pattern of differences between the IRT and CTT short forms point to disparate underlying criteria. For the CTT short form, the criterion is to preserve fidelity by selecting items with the highest factor loadings to provide maximal reliability. By contrast, the IRT criterion is to preserve bandwidth even at the cost of fidelity. Consequently, reliability suffers, but the IRT short form does provide broad coverage of the underlying construct. Nevertheless, IRT approaches can be used to select a short form that preserves fidelity. Achieving this would require that one select items that maximize information at some local value of the underlying construct. For instance, given the data in Table 5, items RC1, RC5, and RC6 can be selected to obtain maximal fidelity around the (from 0 to <1) range of the RC construct. Thus, IRT reveals that scale efficiency often exacts a price, forcing researchers to make a tradeoff between preserving bandwidth and fidelity. The CTT approaches do not reveal such insights.

Table 6
Scale efficiency: correlations and statistics for RC and RA short forms based on CTT and IRT approaches

Construct	Intercorrelations					
	RA(F)	RA(CTT)	RA(IRT)	RC(F)	RC(CTT)	RC(IRT)
RA(F)[a]	1.00					
RA(CTT)[b]	.93	1.00				
RA(IRT)[c]	.93	.84	1.00			
RC(F)[d]	.50	.50	.44	1.00		
RC(CTT)[e]	.48	.46	.41	.90	1.00	
RC(IRT)[f]	.42	.43	.37	.89	.67	1.00
α[g]	.86	.88	.69	.85	.80	.60
Range (ρ)[h]	.20−.76	.64−.76	.32−.58	.24−.72	.48−.72	.26−.46

[a] This composite is based on the full scale of RA items from the Rizzo et al.'s construct.

[b] This composite is based on a "short form" of RA scale selected using CTT procedures {RA4, RA5, and RA6}.

[c] This composite is based on a "short form" of RA scale selected using IRT procedures {RA7, RA4, and RA1}.

[d] This composite is based on the full scale of RC items from the Rizzo et al.'s construct.

[e] This composite is based on a "short form" of RC scale selected using CTT procedures {RC2, RC4, and RC6}.

[f] This composite is based on a "short form" of RA scale selected using IRT procedures {RC7, RC1, and RC3}.

[g] This is the estimated Cronbach's α reliability for the corresponding composite.

[h] This is the range of interitem correlations for the items included in the corresponding composite.

Moreover, IRT provides a unique approach for achieving scale efficiency based on adaptive survey designs (Balasubramaniam and Kamakura, 1989; Singh et al., 1990). Adaptive survey designs are based on the notion that, for any individual, (1) items that do not provide information around the neighborhood of the individual's standing on the underlying construct are not very useful to administer and (2) if the survey could be tailored to administer only those items that are informative, little loss of measurement fidelity will be likely to occur. When surveys are computerized, it is possible to implement such adaptive designs and tailor to each and every individual. Balasubramaniam and Kamakura (1989) and Singh et al. (1990) provided evidence that such adaptive designs obtain measurement precision with fewer items administered to each individual. Such adaptive designs have yet to gain widespread acceptance but IRT-based designs promise to deliver innovative solutions to scale efficiency problems.

6.4. Scale refinement problem

The need for refining the RC and RA constructs is abundantly clear. Rizzo et al.'s constructs, developed almost 30 years ago, and arguably the most popular scales for assessing RC and RA, lack discriminant validity and reliably measure a relatively small bandwidth, when the underlying concepts are in fact rich, comprehensive, and distinct (King and King, 1990). In such situations, researchers face an interesting dilemma: they can either (1) choose to work with what they suspect are deficient constructs or they can (2) start anew by developing a "new" construct for measuring the focal concepts. The conventional wisdom is that it is best to leave the original items intact and use them without major revisions. This "take it or leave it" guideline is probably based on the fear that adding or altering items on a "well-accepted" scale may irrevocably change the meaning of the underlying construct in ways that cannot be easily discerned or determined. At the same time, starting anew poses significant effort and data demands while failing to take full advantage of past scale development work. In addition, this approach promotes proliferation of constructs measuring the same concept, with labeling or procedural differences (e.g., tapping different aspects).

For marketing researchers seeking to deal with the preceding dilemma, the CTT approach offers little guidance. Psychometric indicators such as reliability and validity are properties of an entire scale, making researchers wary of changing such indicators by adding or altering scale items. Thus, it is not surprising that the Rizzo et al.'s RC and RA items have been used for almost 30 years without changes or alterations, regardless of contextual and temporal differences. In this sense, the CTT approach promotes a "take it or leave it" attitude toward well-accepted constructs.

The IRT approach provides a different perspective on the scale refinement dilemma. First, the IRT approach favors *item banking*, whereby additional items are continuously developed and added into a "bank" of items for a given construct. Such a bank can include items that demonstrably yield significant information at different points along the latent continuum (and thus are not redundant) and that collectively widen the bandwidth of measurement. For instance, in the case of RA, the IRT analysis suggests that the item bank is deficient and more items need to be developed that provide information below the mean level (<0 along the latent continuum). Because item characteristics are unaffected by including additional items (as noted in Bejar's test), the IRT approach encourages building construct item banks by identifying bandwidth or fidelity gaps and promoting scale enhancement efforts to address these gaps. Second, the IRT approach favors *item value analysis*, whereby current items are continuously evaluated for their value in providing information about the underlying construct. For instance, in the case of RC, the IRT analysis suggests that item RC5 may offer little value because it (1) has low fidelity (*a* parameter estimate <1), (2) provides peak information around the midrange (0 to <1 on the latent continuum), where several other RC items are effective as well (e.g., RC1), and (3) is likely redundant with items RC1, RC2, and RC6. Of course, the value of RC5 needs to be evaluated in other contexts and samples before deciding to either delete it from the bank or refine its wording/content so that it delivers greater value. In this sense, the IRT approach promotes regular scale refinement through item value analysis and, consequently, helps improve the quality of the construct item bank. Third, the IRT approach allows for *targeted item selection*, whereby researchers can select specific subsets of items from the bank that best fit the purposes of a given study. That is, because it recognizes bandwidth–fidelity tradeoffs, the IRT approach implies that, depending on the purpose of a study, certain items in the bank may not contribute to measurement fidelity. For instance, a researcher interested in identifying salespeople who perceive high levels of RA, perhaps as a screening for a counseling or skill training program, would likely want to use RA items that provide most of their information at the high end of the RA continuum and will be less concerned about fidelity at other points on the latent continuum. In this case, the researcher's purpose is best served by a *peaked* IF for the RA construct, with the peak targeted in the range of, say, >1.5. Consequently, my IRT analysis suggests that in the current RA bank, items RA1, RA3, and RA4 are best suited to the researcher's purpose, and using this subset is unlikely to compromise the desired measurement fidelity of the RA construct.

Taken together, these insights from the IRT approach provide guidelines and encouragement for dealing with the scale refinement dilemma by (1) continuously building item banks by adding items that fill critical bandwidth–fidelity gaps, (2) regularly evaluating the value of individual items to improve the quality of the item bank, and (3) selecting different item subsets depending on the purpose at hand. This "build it, refine it, and use it" approach toward well-established constructs stands in direct contrast to the CTT approach of "take it or leave it" and offers different perspectives on scale refinement efforts (see Table 4 for a summary).

Nevertheless, the IRT approach presented here should be viewed as an introduction to the large variety of IRT models available for different types of data and problems. I utilized several simplifying procedures, including (1) dichotomization of graded response scale, (2) a test of dimensionality based on simple, graphical procedures, and (3) a 2PL IRT model that assumes a monotonic relationship between the underlying latent variable and response probabilities. Approaches for directly analyzing graded response scales (Samejima, 1969) and alternative unfolding IRT models (Roberts et al., 2000) are available but involve greater complexity. Some researchers contend that unfolding IRT models may be more appropriate for binary data, with the choices depending on the specific locations of items and respondents on the latent continuum. Likewise, advances in testing for the unidimensionality of items (Zhang and Stout, 1999; McDonald, 1981) and the availability of multidimensional IRT models (McDonald, 1999; Reckase, 1997) open new avenues for future development and applied research. My aim has been to demonstrate the potential and principles of IRT, and I hope that readers will be encouraged to delve into the IRT literature and select the specific IRT model that is suited to their data and purpose.

7. Concluding comments

> IRT is ultimately a theory about processes underlying a person's response to a question...[IRT] models have been developed in the context of the long history of psychological theory about the processes involved when people answer questions. As a result, the application of these models provides a form of data analysis that may be particularly informative.
>
> Thissen and Steinberg (1988, p. 385).

Theory development in marketing research depends, in part, on better measurement of concepts. The purpose of theory is to explain and predict some phenomenon; the purpose of measurement is to understand the phenomenon *itself*. Thus, progress in measurement logically precedes progress in substantive theory. To facilitate progress in measurement, in this paper, I set out to provide marketing researchers an alternative perspective—one based on IRT—for addressing contemporary measurement concerns. In the tradition of Davis (1971), I have attempted to demonstrate that IRT offers an alternative and *interesting* perspective because it is (1) functionally interesting, revealing that a phenomenon commonly believed to function effectively (i.e., maximizing Cronbach's α) actually functions ineffectively from another standpoint (i.e., that of bandwidth), (2) generalizationally interesting, since it demonstrates that an apparently general phenomenon (e.g., reliability) is in reality a locally defined phenomenon (e.g., information), and (3) assumptionally interesting, as it reveals that some commonly held assumptions (e.g., linear relationships) are in reality not defensible. Nevertheless, it is obvious that IRT-based methods are more complex, lack the inherent simplicity of CTT methods, and require involved computer programs. In addition, IRT models are not a panacea for all measurement woes. Like any other model, IRT imposes its own constraints and limitations. Thus, it is *not* recommended that one use IRT where CTT-based methods are appropriate. Rather, it is important to use the simplest *but most appropriate* measurement model. There seems to be considerable published evidence at this point that CTT assumptions are relatively strict and difficult to meet in most social sciences research (Reise and Widaman, 1999; Reise et al., 1993). In contrast, IRT represents a nonlinear probabilistic model that seems to be consistent with data in several different situations (van der Linden and Hambleton, 1999; Drasgow and Hulin, 1990). Further, IRT-based concepts of information, bandwidth, and fidelity hold promise for understanding concept measurement in significantly greater depth. Finally, as the headnote to this section suggests, IRT is consistent with a cognitive theory of how people respond to questions. For these reasons, I argue that IRT warrants the serious attention of marketing researchers. Future empirical applications of IRT can critically evaluate its significance and contribution to marketing theory and research.

Acknowledgements

This research was partially developed when the author was visiting at the Norwegian School of Management, Bergen to participate in their methodology seminars. The author sincerely acknowledges the helpful comments from the participants of the methodology seminars, Bruce Alford, co-Editor of this special issue, and two anonymous reviewers.

Appendix A. RC and RA items used in the study

RC items	
RC1	I have to do things that should be done differently.
RC2	I receive an assignment without the manpower to complete it.
RC3	I have to buck a rule or policy in order to carry out an assignment.
RC4	I receive incompatible requests from two or more people.
RC5	I do things that are apt to be accepted by one person and not accepted by others.
RC6	I receive an assignment without the adequate resources and materials to execute it.
RC7	I work with two or more groups (people) who operate quite differently.
RC8	I sometimes work on unnecessary things.
RA items	
RA1	I feel certain about how much authority I have. *
RA2	Clear planned goals/objectives exist for my job. *
RA3	I know that I have divided my time properly. *
RA4	I know what my responsibilities are. *
RA5	I know exactly what is expected of me. *
RA6	Explanation is clear for what has to be done. *
RA7	I know how my performance is going to be evaluated. *

* This item is reverse scored.

Appendix B. Mathematical formulations for CTT and IRT

B.1. Mathematical formulation of CTT

In CTT, the observed score X is related to the true score T as follows:

$$X_{ij} = T_{ij} + \varepsilon_{ij} \tag{A2.1}$$

where i indexes the item, j indexes the individual respondent, and ε represents the random error component. The underlying assumptions of CTT include

$$E(X_{ij}) = T_{ij} \tag{A2.2}$$

and

$$\sigma_{\varepsilon T} = 0 \tag{A2.3}$$

In other words, the expected value of the observed response is the individual's true score, and the true score is uncorrelated with the random error (Lord and Novick, 1968). From Eqs. (A2.1), (A2.2), and (A2.3), it can be shown that

$$\sigma_X^2 = \sigma_T^2 + \sigma_e^2 \tag{A2.4}$$

That is, the variance of observed scores is the sum of the variances of true scores and random error. From Eq. (A2.4), reliability is defined as the ratio of $\sigma_T{}^2$ and $\sigma_X{}^2$, which is equal to $(1-\sigma_e{}^2)/\sigma_X{}^2$.

Although the preceding model is statistical in nature (because of ε), the relationship between X_{ij} and T_{ij} is posited to be direct (i.e., not probabilistic).

B.2. Mathematical formulation of IRT

B.2.1. Equations

The basic equation in IRT defines a nonlinear response function that relates the location of an individual *j* on the underlying trait (θ) to the probability of *j*'s agreeing with a specific item *i*.

$$P_i(\theta) = \frac{1}{1+\exp[-a_i(\theta_j - b_i)]} \tag{A2.5}$$

where $P_i(\theta)$ = probability of agreeing with item *i*, a_i = discrimination or sensitivity parameter for item *i*, b_i = threshold or affectivity parameter for item *i*, θ_j = standing on an underlying latent trait for respondent *j*.

Let V=$(k_1,k_2,\ldots,k_n)$ be the observed response pattern of the *j*th respondent (j = 1,2,. . ., N) to the *n* items. Each k_i, for instance, can either be 0 (disagree) or 1 (agree). Let $P_V(\theta)$ be the probability of response pattern *V* given that the respondent *j* has θ_j trait level. That is,

$$P_V(\theta) = \text{Prob.}[(k_1,k_2,\ldots k_n)/\theta_j] \tag{A2.6}$$

IRT models utilize the principle of local independence in order to simplify the Eq. (A2.5). The notion of local independence implies that the variances and covariances among item responses can be attributed to only three sources: (a) systematic variance that can be accounted for by an underlying latent trait θ, (b) unique variance specific to the item, and (c) random error variance. Thus, under local independence, Eq. (A2.6) becomes:

$$P_V(\theta) = \text{Prob.}(k_1/\theta_j),\text{Prob.}(k_2/\theta_j)\ldots\text{Prob.}(k_n/\theta_j) \tag{A2.7}$$

In Eq. (A2.7), the probability of the joint distribution of responses is written as the product of the conditional probabilities for individual items. Noting that Prob.(k_i) equals $P_i(\theta)$ if k_i = 1 and $Q_i(\theta)$=$[1-P_i(\theta)]$ if otherwise (i.e., k_i = 0), Eq. (A2.7) becomes Eq. (A2.8):

$$\begin{aligned} P_V(\theta) &= P_{k_1}(\theta)^{k_1}Q_{k_1}(\theta)^{1-k_1} * P_{k_2}(\theta)^{k_2}Q_{k_2}(\theta)^{1-k_2 *}\ldots \\ &\quad *P_{k_n}(\theta)^{k_n}Q_{k_n}(\theta)^{1-k_n} \\ &= \sum_{i=1}^{n} P_{k_i}(\theta)^{k_i}Q_{k_i}(\theta)^{1-k_i}.N(\theta).\text{d}(\theta) \end{aligned} \tag{A2.8}$$

The marginal probability for obtaining a response pattern *V* is obtained by integrating over *N*(θ) the distribution of θ in the population of interest.

$$P_V = \int_{-\infty}^{+\infty} \sum P_{k_i}(\theta)^{k_i}Q_{k_i}(\theta)^{1-k_i}.N(\theta).\text{d}\theta \tag{A2.9}$$

B.2.2. Estimation issues

Several methods for parameter estimation are now available, such as the conditional maximum likelihood (CML), joint maximum likelihood (JML), and MML methods (e.g., see Hambleton and Swaminathan, 1985). In general, both the item (i.e., a_i and b_i) and trait (i.e., θ) parameters are unknown. However, under the assumptions of random regressors, the latent trait θ is assumed to be a random variable and with appropriate distributional assumptions regarding θ (usually normality), the trait estimates are removed from the estimation of item parameters by integrating them out of the likelihood function. This approach is consistent with the common factor model (McDonald, 1982). In addition, for "calibration" samples, to which the IRT model may be initially fit, the trait estimates are rarely of interest in and of themselves. Rather, item parameters are required from the calibration sample in order to estimate the location of future respondents on the underlying latent construct.

Thissen (1982) and Bock and Aitkin (1981) have proposed an effective and efficient approach for the estimation of item parameters under the random regressors model. This approach is based on the maximization of the log-likelihood function for the marginal probability distribution of *V* (Eq. (A2.9)). Hence, this is often referred to as the MML method. Bock and Aitkin (1981) show that the EM algorithm can be modified to implement this approach. In particular, marginal estimators for item parameters are obtained by iteratively repeating the EM steps until the process converges to the following criterion (Eq. (A2.10)):

$$\text{Maximize}: \quad \log L(a_i, b_{ik}) = \log\sum \int \sum P_{k_i}(\theta)^{k_i}Q_{k_i}(\theta)^{1-k_i}.N(\theta).\text{d}(\theta) \tag{A2.10}$$

Statistical properties of the MML estimators for IRT models have not yet been conclusively established. In contrast to other available procedures (e.g., JML), however, the marginal estimators have an "important advantage" because of their "theoretical accuracy" (Lord 1980, p. 158). Hambleton and Swaminathan (1985) suggested that such estimators might have desirable attributes such as consistency and asymptotic normality. In addition, the MML method offers significant advantages when the number of scale items is "relatively" small (e.g., 10 or less). Several computer programs are now available to estimate IRT parameters (e.g., LOGOG and BILOG). Thissen's (1991) MULTILOG is particularly flexible as well as comprehensive for MML estimators of IRT models.

Appendix C. MULTILOG program lines for estimating item parameters of RC and RA items

```
>TITLE
ESTIMATING MML ITEM PARAMETERS FOR ROLE
   CONFLICT AND ROLE AMBIGUITY ITEMS FOR
   SME DATA
>PROBLEM RANDOM NITEMS=15 NGROUPS=1
   INDIVIDUAL NEXAMINEES=472;
>TEST ALL L2
>ESTIMATE NCYCLES=100 BIG;
>END;
HOW MANY RESPONSE CODES IN RAW DATA?
2
ENTER CODES 2A1
01
ENTER VECTOR OF CORRECT RESPONSES, 79A1
111111111111111
IS ANY CODE MISSING OR NOT-REACHED?
   (Y OR N)
N
ENTER FORMAT FOR DATA
(15A1, 18X, F2.0)
```

References

Andrich D. A rating formulation for ordered response categories. Psychometrika 1978;43(4):561–73 (Dec.).

Balasubramaniam S, Kamakura W. Measuring consumer attitudes toward the marketplace with tailored interviews. J Mark Res 1989;26: 311–26 (August).

Bearden W, Netemeyer RG. Handbook of marketing scales: multi-item measures for marketing and consumer behavior research. 2nd ed. Newbury Park, CA: Sage Publications, 1999.

Behrman D, Perreault W. A role stress model of the performance and satisfaction of industrial salespersons. J Mark 1984;48(4):9–21 (Fall).

Bejar II. A procedure for investigating the unidimensionality of achievement tests based on item parameter estimates. J Educ Meas 1980;17(4): 283–96 (Winter).

Biddle BJ. Recent developments in role theory. Annu Rev Sociol 1986;12: 67–92.

Birnbaum A. Test scores, sufficient statistics, and the information structures of tests. In: Lord FM, Novick MR, editors. Statistical theories of mental test scores. Reading, MA: Addison-Wesley Publishing, 1968. p. 425–52.

Blalock HM. The measurement problem in methodology in social research. New York: McGraw-Hill, 1968.

Blalock HM. Basic dilemmas in the social sciences. Beverly Hills, CA: Sage Publications, 1984.

Bock RD. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika 1972;37: 29–51 (March).

Bock RD, Aitkin M. Marginal maximum likelihood estimation of parameters: an application of the EM algorithm. Psychometrika 1981;46: 443–59 (December).

Bollen K, Lennox R. Conventional wisdom on measurement: a structural equations perspective. Psychol Bull 1991;110(2):305–14 (September).

Bradlow ET, Wainer H, Wang X. A Bayesian random effects model for testlets. Psychometrika 1999;64(2):153–68 (June).

Campbell JP, Daft RL, Hulin CL. What to study: generating and developing research questions. Beverly Hills, CA: Sage Publications, 1982.

Churchill GA. A paradigm for developing better measures of marketing constructs. J Mark Res 1979;16(1):64–73 (February).

Churchill GA, Peter JP. Research design effects on the reliability of rating scales. J Mark Res 1984;21:360–75 (November).

Churchill GA, Ford N, Hartley S, Walker O. The determinants of salesperson performance: a meta-analysis. J Mark Res 1985;22:103–18 (May).

Cohen J. The cost of dichotomization. Appl Psychol Meas 1983;7(3): 249–53 (Summer).

Davis MS. That's interesting! Toward a phenomenology of sociology and a sociology of phenomenology. Philos Soc Sci 1971;309–44.

Drasgow F, Hulin C. Item response theory. In: Dunnette M, Hough L, editors. Handbook of industrial and organizational psychology. Palo Alto, CA: Consulting Psychologists Press, 1990. p. 577–636.

Drasgow F, Parsons CK. Application of unidimensional item response theory models to multidimensional data. Appl Psychol Meas 1983; 7(2):189–99.

Edwards J. A cybernetic theory of stress, coping, and well-being in organizations. Acad Manage Rev 1992;17(2):238–74 (April).

Fischer G. Some neglected problems in IRT. Psychometrika 1995;60(4): 459–87 (December).

Fisher C, Gitelson R. A meta-analysis of the correlates of role conflict and role ambiguity. J Appl Psychol 1983;68(2):320–33 (May).

Ford N, Walker OC, Churchill G. Expectation-specific measures of intersender conflict and role ambiguity experienced by industrial salesman. J Bus Res 1975;3(2):95–110 (April).

Fry LW, Futrell CM, Parasuraman A, Chmielewski M. An analysis of alternative causal models of salesperson role perceptions and work-related attitudes. J Mark Res 1986;23(2):153–63 (May).

Gaito J. Measurement scales and statistics: resurgence of an old misconception. Psychol Bull 1980;87(3):564–7 (May).

Gerbing DW, Anderson JC. An updated paradigm for scale development incorporating unidimensionality and its assessment. J Mark Res 1988; 25(2):311–26 (May).

Green BF, Bock RD, Humphreys LG, Linn RL, Reckase MD. Technical guidelines for assessing computerized adaptive tests. J Educ Meas 1984;21:347–60.

Greer SA. The logic of social inquiry. Chicago, IL: Aldine Pub., 1969.

Hambleton RK, Swaminathan H. Item response theory: principles and applications. Boston, MA: Kliweer-Nijhoff, 1985.

Hambleton RK, van der Linden WJ. Advances in item response theory and applications—an introduction. Appl Psychol Meas 1982;6(4): 373–8 (Fall).

Horn JL. A rationale and test for the number of factors in factor analysis. Psychometrika 1965;30:179–85.

House RJ, Schuler RS, Levanoni E. Role conflict and ambiguity scales: reality or artifacts. J Appl Psychol 1983;68(2):334–7 (May).

Howell RD, Wilcox JB, Bellenger DN, Chonko LB. An assessment of the role conflict and role ambiguity scales. AMA Educ Proc 1988;54: 314–9.

Hulin CF, Drasgow F, Parsons C. Item response theory. Applications to psychological measurement. Homewood, IL: Dow-Jones-Irwin, 1983.

Humphreys LG, Montanelli RG. An investigation of the parallel analysis criterion for determining the number of common factors. Multivariate Behav Res 1975;10(2):193–205 (April).

Idaszak J, Drasgow F. A revision of the job diagnostic survey: elimination of a measurement artifact. J Appl Psychol 1987;72(1):69–74 (February).

Jackson S, Schuler R. A meta-analysis and conceptual critique of research on role ambiguity and role conflict in work settings. Organ Behav Hum Decis Processes 1985;36(1):16–78 (August).

Kahn RL, Wolfe DM, Quinn RP, Snoek JD. Organizational stress: studies in role conflict and ambiguity. New York, NY: Wiley, 1964.

King LA, King DW. Role conflict and role ambiguity: a critical assessment of construct validity. Psychol Bull 1990;107(1):48–64 (January).

Lewis C. Test theory and psychometrika: the past twenty-five years. Psychometrika 1986;51(1):11–22 (March).

Lord FM. A theory of test scores. Psychometric Monogr 1952;7.

Lord FM. Applications of item response theory to practical testing problems. New Jersey: Lawrence Erlbaum, 1980.

Lord FM, Novick MR. Statistical theories of mental test scores. Reading, MA: Addison-Wesley Publishing, 1968.

Marsh H, Balla J, Hau K-T. An evaluation of incremental fit indices: a clarification of mathematical and empirical properties. In: Marcoulides G, Schumacker R, editors. Advanced structural equation modeling: issues and techniques. Mahwah, NJ: Erlbaum Associates, 1996. p. 315–45.

Masters GN. A Rasch model for partial credit scoring. Psychometrika 1982;47(2):149–74 (June).

McDonald RP. The dimensionality of tests and items. Br J Math Stat Psychol 1981;34:100–17 (May).

McDonald RP. Test theory: a unified approach Mahwah, NJ: LEA Publishers, 1999.

McGee GW, Ferguson CE, Seers A. Role conflict and role ambiguity: do the scales measure these two constructs? J Appl Psychol 1989;74(5): 815–8 (October).

Mellenbergh GJ. Measurement precision in test score and item response models. Psychol Methods 1996;1(3):293–9 (September).

Michaels RE, Day RL, Joachimsthaler EA. Role stress among industrial buyers: an integrative model. J Mark 1987;51(2):28–45 (April).

Michell J. Measurement scales and statistics: a clash of paradigms. Psychol Bull 1986;100(3):398–407 (November).

Mokken RJ, Lewis C. A nonparametric approach to the analysis of dichotomous item responses. Appl Psychol Meas 1982;6(4):283–98 (Fall).

Nunnally J. Psychometric theory. 2nd ed. New York: McGraw-Hill, 1978.

Pearce JL. Bringing some clarity to role ambiguity research. Acad Manage Rev 1981;6(4):665–74 (October).

Rasch G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institute for Educational Research, 1960.

Reckase MD. A linear logistic multidimensional model for dichotomous item response data. In: van der Linden WJ, Hambleton R, editors. Handbook of modern item response theory. New York: Springer-Verlag, 1997. p. 278–86.

Reise S, Widaman KF. Assessing the fit of measurement models at the individual level: a comparison of item response theory and covariance structure approaches. Psychol Methods 1999;4(1):3–21 (March).

Reise SP, Widaman KF, Pugh RH. Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance. Psychol Bull 1993;114(3):552–6 (November).

Reiser M. Latent trait modeling of attitude items. In: Bohrnstadt G, Borgatta E, editors. Social measurement. Thousand Oaks, CA: Sage Publications, 1981. p. 147–61.

Richardson MW. The relationship between difficulty and the differential validity of a test. Psychometrika 1936;1:33–49.

Rizzo JR, House R, Lirtzman SI. Role conflict and ambiguity in complex organizations. Adm Sci Q 1970;15:150–63.

Roberts JS, Donoghue JR, Laughlin JE. A general item response theory model for unfolding unidimensional polytomous responses. Appl Psychol Meas 2000;24(1):3–32 (March).

Samejima F. Estimation of latent ability using a response pattern of graded scores. Psychometrika Monogr 1969;4:1–100 (Part 2).

Samejima F. Weakly parallel tests in latent trait theory with some criticisms of classical test theory. Psychometrika 1977;42(2):193–9 (June).

Schmitt N, Stults DM. Factors defined by negatively keyed items: the result of careless respondents? Appl Psychol Meas 1985;9(4): 367–73 (December).

Schuler RS. Role perceptions, satisfaction and performance moderated by organizational level and participation in decision making. Acad Manage J 1977;20(1):159–65 (March).

Schwab DP. Construct validity in organizational behavior. In: Staw BM, Cummings LL, editors. Research in organizational behavior, vol. 2. Greenwich: JAI Press, 1980. p. 2–43.

Singh J. Boundary role ambiguity: facets, determinants, and impacts. J Mark 1993;57(2):11–31 (April).

Singh J, Rhoads G. Boundary role ambiguity in marketing oriented positions: a multidimensional, multifaceted operationalization. J Mark Res 1991;28(3):328–38 (August).

Singh J, Howell R, Rhoads G. Adaptive designs for Likert-type data: an approach for implementing marketing surveys. J Mark Res 1990; 27(3):304–21 (August).

Singh J, Goolsby J, Rhoads G. Behavioral and psychological consequences of boundary spanning burnout for customer service representatives. J Mark Res 1994;31(4):558–69 (November).

Stevens SS. On the theory of scales of measurement. Science 1946;103: 667–80.

Thissen D. Marginal maximum likelihood estimation for the one parameter logistic model. Psychometrika 1982;47(2):175–86 (June).

Thissen D. MULTILOG. Indiana: Scientific Software, 1991.

Thissen D, Steinberg L. A taxonomy of item response models. Psychometrika 1986;51(4):567–78 (December).

Thissen D, Steinberg L. Data analysis using item response theory. Psychol Bull 1988;104(3):385–95 (November).

Tracy L, Johnson TW. What do the role conflict and role ambiguity scales measure? J Appl Psychol 1981;66(4):464–9 (August).

van der Linden WJ, Hambleton RK. Handbook of modern item response theory. New York: Springer-Verlag, 1997.

Van Schuur WH, Kiers HAL. Why factor analysis is often the incorrect model for analyzing bipolar concepts, and what model can be used instead? Appl Psychol Meas 1994;18:97–110.

Van Sell M, Brief AP, Schuler RS. Role conflict and role ambiguity: integration of the literature and directions for future research. Hum Relat 1981;34(1):43–71 (January).

Weiss DJ, Davison ML. Test theory and methods. Annu Rev Psychol 1981;32:629–58.

Whetten DA. Coping with incompatible expectations: an integrated view of role conflict. Adm Sci Q 1978;23(2):254–71 (June).

Wright BD, Stone MH. Best test design: Rasch measurement. Chicago, IL: MESA Press, 1977.

Zeller RA, Carmines EG. Measurement in the social sciences. Cambridge, MA: Cambridge Univ. Press, 1980.

Zhang J, Stout W. The theoretical DETECT index of dimensionality and its application to approximate simple structure. Psychometrika 1999;64(2): 213–50 (June).