AN ALTERNATE APPROACH TO ASSESSING CROSS-CULTURAL MEASUREMENT EQUIVALENCE IN ADVERTISING RESEARCH

Michael T. Ewing, Thomas Salzberger, and Rudolf R. Sinkovics

ABSTRACT: This paper offers a new methodological framework to guide researchers attempting to quantitatively assess how a pluralistic audience perceives a standardized television advertisement. Rasch (1960) measurement theory is introduced as an alternative to the more commonly employed multigroup confirmatory factor analysis (CFA) approach to assessing cross-cultural scalar equivalence. By analyzing a multicultural data set, we are able to make various inferences concerning the scalar equivalence of Schlinger's confusion scale. The methodology reveals the limits of the scale, which in all probability would not have been detected using traditional approaches. For researchers attempting to develop new scales, or even to refine existing scales, strict adherence to established guidelines of item generation together with the application of the proposed methodology should ensure better results for both theorists and practitioners.

There is considerable evidence to suggest that attitude toward the advertisement (A_{ad}) influences brand attitudes in pretest situations and in understanding, predicting, and perhaps even forestalling wear-out (Lutz 1985). In this regard, reaction profiles have been found to predict A_{ad} both directly (Burke and Edell 1982) and indirectly via ad perceptions (Lutz 1985). Reaction profiles yield more easily quantifiable data and are therefore more amenable to repeated routine use. Moreover, an inherent advantage of closed-ended responses is ease of comparison of findings across ads and/or studies (Lutz 1985). Reaction profiles can also be used in a diagnostic way, either to assess how an advertisement is perceived on various dimensions and/or to better understand its total impact on the audience (Aaker and Stayman 1990). But what if "the audience" is culturally diverse? As Andrews, Durvasula, and Netemeyer (1994) note, examining the cross-national applicability of advertising measures has become increasingly important. As Taylor points out, however, "too often, in past . . . research, authors have not used available statistical techniques for ensuring the equivalence of data collected cross-culturally" (2003, p. 247). Hence, this paper considers the use of reaction profiles to test standardized advertisements in multicultural environments.

This study differs from previous research in two significant ways, however. First, rather than reexamine measurement equivalence cross-nationally, we have taken up Lenartowicz, Johnson, and White's (2003) recent call to first account for intracountry cultural variation to avoid erroneous nonsignificant findings of cross-national differences by disregarding cultural variations within a country. Second, instead of relying solely on a multigroup confirmatory factor analysis (CFA) approach to assessing scalar equivalence, which has been widely used in marketing and international business generally (cf. Knight, Spreng, and Yaprak 2003; Steenkamp and Baumgartner 1998) and advertising specifically (cf. Andrews, Durvasula, and Netemeyer 1994; Durvasula et al. 1993; Ewing, Caruana, and Teo 2002; Ewing, Caruana, and Zinkhan 2002), we introduce an alternative measurement approach based on Rasch (1960) theory. We hope that this novel application will stimulate increased interest and debate concerning cross-cultural testing and measurement in advertising. It is important to note that we do not advocate one approach over the other. Rather, we challenge future researchers to continue along this line of inquiry and draw appropriate conclusions themselves.

PROBLEM AND PURPOSE

A striking anomaly in today's international advertising environment is the parallel evolution of both global and

Michael T. Ewing (D.Com., University of Pretoria) is a professor of marketing in the Faculty of Business and Economics, Monash University, Australia.

Thomas Salzberger (Ph.D., Vienna University of Economics and Business Administration) is an assistant professor of marketing, Vienna University of Economics and Business Administration, Austria.

Rudolf R. Sinkovics (Ph.D., Vienna University of Economics and Business Administration) is a lecturer, University of Manchester, Manchester Business School, United Kingdom.

The authors contributed equally; names appear in alphabetical order. The authors sincerely thank Erik du Plessis, CEO of Millward Brown Impact (South Africa), for access to proprietary data. They also thank Ray Taylor and four anonymous reviewers for constructive comments on previous drafts of this manuscript.

Journal of Advertising, vol. 34, no. 1 (Spring 2005), pp. 17–36.

ISSN 0091-3367 / 2005 $9.50 + 0.00.

micromarketing theory and practice. Markets are getting simultaneously bigger and smaller (de Mooij 1994). Coca-Cola is an excellent example of this paradox. The world's most ubiquitous brand has adopted a "think local, act local" strategy (James 2001). Against this backdrop, rather than focus on the commonly considered case of cross-national variation, we begin by examining the underresearched issue of intracountry variation (Lenartowicz, Johnson, and White 2003), first highlighted as a salient issue within an advertising context by Deshpande and Stayman (1994). So, for example, should the same television advertisement flighted in Maine or Utah be used in Florida? Can the same advertisement appeal equally to African Americans, Chinese Americans, Hispanic Americans, and other population groups? Clearly, the standardization–adaptation issue applies in both inter- and intracountry contexts. While our empirical focus is on the latter, our approach is equally generalizable and applicable to both.

In the ensuing study, we subscribe to Onkvisit and Shaw's (1987) definition of standardization, with one minor change: "an advertisement which is used [cross-culturally] with virtually no change in its theme, copy, or illustration, except for translation when needed."

Revisiting the Case for Advertising Response Scales

Multiple-item rating scales seek to capture an individual's immediate reaction to an advertisement and to understand how the ad works from the consumer's perspective. They allow researchers to capture a wide range of cognitive and emotional reactions to an advertisement that would otherwise be lost or captured incorrectly if the respondent were to try to describe these feelings either verbally or in writing (Zinkhan and Fornell 1985). The primary advantage of response scales is that they provide a stable list of items that can be used to track reactions to an advertisement (Zinkhan and Burton 1989). It is perhaps not surprising that reaction profiling was pioneered by advertising practitioners and that most research using this approach has been conducted by industry (Lutz 1985). Indeed, the data analyzed in this paper was supplied by industry (www.impact.co.za) and is based on Schlinger's (1979) well-known Viewer Response Profile (VRP). However, despite the many reaction profiles that have been developed over the years (16 between 1964 and 1983 alone), there has always been a lack of consensus with respect to their underlying dimensionality (Lutz 1985). Given that our particular focus is cross-cultural measurement equivalence, for illustrative (and practical) purposes, we will only analyze one Schlinger VRP dimension—the four-item confusion scale. Of course, our method is not restricted to response profiles. It could be used to equal effect with ad-execution cognitive responses or any of the other cognitive or affective antecedents of A_{ad} (Lutz 1985).

Advertising Within a South African Context

South Africa provides an ideal setting for conducting cross-cultural advertising research. The country contrasts a striking blend of developed and developing-world business, social, and cultural environments. It is a truly multicultural market—hence the popular moniker, "rainbow nation."

Income distribution in South Africa varies from first-world wealth to third-world poverty, most of the world's major religions are represented in significant numbers, and the country boasts 11 official languages. Whereas the present government, in power since 1994, has attempted to unify racial and cultural groups, prior to that, the minority white Apartheid administration attempted to enforce cultural and racial differences by means of policy and legislation. In many instances, these differences still persist, perhaps as a legacy of apartheid, but possibly because many cultural differences are deep-seated. Hofstede's (1984) well-known study of cultural differences among nations found [his predominantly white male] South African respondents to rate high on individualism, while the black African cultures tend to value and reward collectivism (cf. Mbigi 2000). South Africa possesses a relatively sophisticated broadcast media infrastructure, and a well-established, highly successful advertising industry. Indeed, local agencies have a remarkable track record in winning many of the world's premier advertising awards (see, e.g., www.huntlas.co.za), to an extent far outweighed by both total national and per capita advertising spending. Therefore, South Africa provides a highly suitable venue for conducting cross-cultural advertising research: In many ways it is the world in one nation.

CONCEPTUAL BACKGROUND

The standard approach to test a scale's equivalence across different populations is based on multigroup CFA (Steenkamp and Baumgartner 1998). When combined with qualitative reasoning and consideration of item content, this methodology is a suitable approach within the standard measurement framework in the social sciences. Issues surrounding manifest and latent variables and distributional assumptions can be problematic, although astute researchers are generally cognizant of these problems. The CFA approach does not overcome many of the limitations of the underlying definition of measurement, however. We therefore consider an alternate methodology—Rasch—which has already gained considerable grounds in educational research, psychology, and medicine (see, e.g., Bond and Fox 2001; Tesio 2003; Wright and Masters 1982).

Given that "interdisciplinary exchange is not keeping up with progress within disciplines" (Taylor 2003, p. 246), this is an attempt to contribute to advertising theory by drawing on work that originated in other disciplines. Salzberger, Sinkovics, and Schlegelmilch (1999) provide an introduction to Rasch measurement theory (Rasch 1960) as a way to assess scalar equivalence. When comparing this approach with the multigroup CFA methodology, Salzberger, Sinkovics, and Schlegelmilch (1999, p. 35) conclude that the Rasch model "offers a superior avenue to measuring latent constructs in general and in cross-cultural settings in particular." The reason lies mainly in the definition of measurement and the special properties of the Rasch model. We will now briefly review how measurement is defined in mainstream marketing/advertising research, before describing the approach in more detail.

Most publications in the social sciences do not clearly elucidate the definition of measurement adopted. Implicitly, Stevens's (1946, 1951) representational view is almost universally followed and accepted. This view holds that measurement is the assignment of numerals or numbers to objects, or, strictly speaking, attributes of objects according to a consistent rule (Stevens 1951). The crucial point is that this definition presumes measurement rather than defines what has to be fulfilled to constitute measurement. The manifest data are immediately seen as being measures.

With his seminal publication, Churchill (1979) established what has almost become the de facto norm for construct measurement in marketing. While this approach has been refined over the past quarter century, the very foundation of measurement is still essentially the same. In the words of Michell (1999), social sciences have concentrated on the "instrumental task" of measurement, that is, devising procedures and instruments to measure latent constructs. Yet we have failed to address the "scientific task" of measurement, namely, whether measurement has been achieved at all. The latter depends not only on the quality of the instrument, but also on the fact that the attribute really is quantitative in nature. Neglecting the scientific task "sounds a warning to us about the nature of measurement in the social sciences" (Balnaves and Caputi 2001, p. 51). If we follow the classical definition of measurement, which prevails in the natural sciences (and there is no reason why we should not), then measurement is concerned with "the discovery or estimation of the ratio of a magnitude of a quantity to a unit of the same quantity" (Michell 1999, p. 222). In this view, numbers are seen as existing independent of measurement, and these numbers can be revealed empirically. Michell (1990) explicates in detail the axioms that have to be satisfied to conclude validly that a variable is really quantitative and that measurement is achieved. In practice, it is a hierarchy of cancellation conditions to which data has to adhere.[1] The Rasch (1960) model is compatible with these axioms of measurement, whereas more complicated models (e.g., the two-parameter logistic model by Birnbaum 1968), belonging to the realm of item response theory (IRT), violate the cancellation conditions. We therefore use the Rasch model as the model of measurement in this study.

After delineating the conceptual and philosophical differences between the standard approach to measurement and the Rasch model, we will now clarify these fundamental issues further by introducing a hypothetical example. All three fictitious items attempt to capture the entertainment value of a television commercial. The items are:

1. The commercial is entertaining.
2. The commercial is a good example of a highly entertaining television commercial.
3. The commercial is one of the most entertaining commercials I have ever seen.

Obviously, these items can be ordered from easily endorsable (*item a*) to most difficult to endorse (*item c*). Consequently, a moderately entertaining ad might provoke respondents to agree to *item a,* but disagree to both *item b* and *item c.* The standard model ignores the feature of "endorsability," however, and focuses instead on the items' intercorrelations. High inter-item correlations are indicative of a scale's quality, but due to floor and ceiling effects, the correlations are maximized with equal endorsability across all items (see Embretson and Reise 2000, p. 36f. for a discussion of so-called difficulty factors produced by factor analysis). In contrast, Rasch modeling investigates the structure of the responses depending on the items' different endorsability. The following pattern (referred to as Guttman pattern) of responses is expected under the hypothesis that the three items are ordered properly.

Item a	Item b	Item c	Raw score
Disagree	Disagree	Disagree	0
Agree	Disagree	Disagree	1
Agree	Agree	Disagree	2
Agree	Agree	Agree	3

Unlike the deterministic Guttman approach (1950), the Rasch model is probabilistic, that is, it accounts for slight deviations due to random factors influencing the response process. Sufficient variation in the items' endorsability is therefore vital to any meaningful Rasch analysis of a scale. In an excellent comparison of the standard model based on classical test theory (CTT) and item response theory, Singh (2004) refers to this issue as the "bandwidth–fidelity problem." The standard approach maximizes fidelity, that is, reliability, by maximizing the inter-item correlations, whereas item response theory, a wider framework of logistic models subsuming the Rasch model, attaches importance to bandwidth, that is, variation in endorsability. What seems like a conflict is, from a Rasch perspective, actually no true tradeoff. The reason is that the reliability of a measure (i.e., technically speaking, the standard error of measurement), depends on the bandwidth of a measurement instrument. In the standard model, the standard error is constant for all levels of measures and is inversely related to reliability (for formula, see Traub

1994). In the Rasch model, the standard error depends on the information that a particular set of items yields for a particular person. Information is maximized when the item's endorsability matches the person's location of the latent dimension (for formula, see Embretson and Reise 2000).

Covering a wide range of the latent dimension is indispensable for measurement for two reasons. First, bandwidth allows for differentiating between response patterns that are expected and those that are not. This is crucial for fulfilling the scientific task of measurement because with all items being equally endorsable, any pattern has the same likelihood. Second, bandwidth guarantees the existence of items that yield enough information for many respondents to ensure a small standard error of measurement. If all items are endorsed too easily (or not easily enough), the targeting of the measurement instrument is potentially suboptimal. In this case, standard errors are large and the model fit cannot be assessed properly. It should be noted that the standard model would not encounter serious problems. In fact, reliability can even increase with such off-target items.

Nevertheless, applications of the Rasch model are still scarce in marketing and advertising research. Soutar and his colleagues (Soutar, Bell, and Wallis 1990; Soutar and Cornish-Ward 1997; Soutar and Ryan 1999) published Rasch analyses of "Consumer Acquisition Patterns for Durable Goods," "Ownership Patterns for Durable Goods and Financial Assets," and of "People's Leisure Activities," respectively. Sinkovics, Salzberger, and Holzmüller (1998) dealt with measurement issues in multicultural research, including the Rasch approach. and Salzberger, Sinkovics, and Schlegelmilch (1999) have compared "Classical Test Theory and Latent Trait Theory Based Approaches." Applications of models that do not belong to the family of Rasch models have been published by Balasubramaniam and Kamakura (1989), Singh, Howell, and Rhoads (1990), and Singh (2004). In a recent review of methodologies for organizational research in management, Schaffer and Riordan (2003) identify two best practices for assessing equivalence. The first is based on covariance structure analysis (i.e., multigroup CFA), the second on item response theory. Hence, we are beginning to witness a slow penetration of alternative measurement models into the wider marketing literature.

Rasch Measurement Theory

The "Rasch model" is actually a "family" constituting a growing number of models (or model variants) tailored to various situations and formats of data. It is the set of features that these models have in common that distinguishes them from other models, such as the Birnbaum model (Birnbaum 1968) or, for polytomous data, the Graded response model (Samejima 1997). In some respects, Rasch models are unique. Many researchers are aware of the Rasch model and of its special properties. However, we argue that its implications for measurement are not widely appreciated, and consequently, applications are scarce.

The Rasch model is sometimes met with resistance because it is seen as being too simple. Clearly, the more parameters a model permits, the better the data can be mirrored. The implicit objective of measurement in this sense is the description of data. But do we really want to describe data? If we want to measure an attribute, then we need to reveal the property of quantity. If we want to reveal the property of quantity, we have to observe the axioms of additive conjoint measurement. Choosing a model that conflicts with these axioms contradicts our initial objective of measurement. It should be noted that the model cannot, and should not, miraculously yield measurement where underlying variables are not quantitative or items are unsuitable to quantify a latent variable. The capacity to best explain data does not lie in the philosophy of the model; rather, the admittedly demanding requirements of measurement are incorporated in the model. The simplicity is a direct consequence of the stringency to which data has to comply to constitute measurement.

It is often argued that it is difficult to achieve measurement in the social sciences, whereas in the natural sciences it is comparatively easier. Indeed, social scientists have to deal with intensive (i.e., not directly observable) attributes that do not reveal their additive structure as conveniently as extensive variables (such as length, for example). Despite the inherent difficulty of measurement in the social sciences, measurement scales abound in marketing and advertising. It is still not clear, however, whether this energetic productivity is cause for pride or critical reflection within the discipline.

The Rasch model offers a different perspective of measurement. We would encourage researchers to view the Rasch model as a parallel approach to measurement. A theoretical comparison of different approaches may help researchers decide which they feel more comfortable with. Assessment of how useful Rasch modeling can actually be in advertising research can only be based on exemplary applications, however.

Special Properties of the Rasch Model

The Rasch model has some unique features worthy of closer examination. An important implication of equal item discrimination is the sufficiency of the unweighted total person (item) raw score statistic for the estimation of the person (item) location parameter. The raw score sufficiency is particularly important because it enables the separate estimation of item and person parameters. Rasch (1961, 1977) referred to this theoretical concept as *specific objectivity.* This means that the estimation of the person parameters is independent of the particular items employed and, vice versa, the estimation of item parameters is independent of the persons used. Specifically, there are no distributional assumptions in terms of the item and/or the person parameters. Another way to express the essence of specific objectivity is the invariance of the parameters. Regardless of the persons used (e.g., samples from different subcultures), the true item parameter is always the same. Empirical evidence to the contrary indicates that the measures are not comparable across the samples.

Dimensionality is another important concern of researchers attempting measurement. While unidimensionality may well be the ultimate goal, theoretically, three cases should be distinguished. First, the construct may be unidimensional, that is, there is only one dimension of interest. In practice, even in this case, there are always other factors that have an impact on the response process. The point is, however, that we do not want to model these additional dimensions, but regard them as a source of error. Hence, a unidimensional measurement model is perfectly reasonable. Second, the construct may be hypothesized to be multidimensional in itself, that is, there are two or more dimensions of interest. Even in this case, one should attempt to separate the dimensions of interest, that is, there are items pertaining to one dimension and other items indicating another. Consequently, the Rasch model can be applied to each dimension separately. It should be noted that in factor analysis we also favor items that can be assigned to one dimension unambiguously, that is, there is no multidimensionality within a particular item, but different items are related to different dimensions. Third, more often than not we hypothesize a set of items to be unidimensional while a multidimensional model would be tenable and theoretically useful. Then, a Rasch analysis of items covering a multitude of dimensions would not be appropriate. Since the prerequisite of unidimensionality is violated in this case, several items should misfit. That is why it is always advisable to carefully scrutinize discarded items to see whether they might have something in common. Then, a subsequent Rasch analysis of initially deleted items might reveal a second dimension "hiding" in the data.

Finally, we want to emphasize that the invariance property as a consequence of specific objectivity and unidimensionality is always subject to data fitting the model. It is in no way provided by the Rasch model automatically and necessarily. Unidimensionality is sometimes mistakenly seen as an unproven assumption of the model. In fact, whether the properties of the model do apply in a particular case fully depends on the fit of the data to the model. If data are inherently multidimensional, the data will not stand all tests of fit. Likewise, if the estimation of item parameters does depend on the persons used, then the invariance property does not hold. In other words, the properties of the model are hypotheses about the data. They have to be raised if we are striving for measurement, and they have to be tested empirically by assessing model fit if we want to check whether we have achieved measurement or not.

The Rasch Model and Cross-Cultural Comparability

The investigation of cross-cultural comparability using the Rasch model capitalizes on the invariance property. Invariance is fundamental for the generalizability of the measurement instrument across cultural borders as well as different sorts of reference objects (specific advertisements in the case of attitude toward advertisements). If the item estimates do depend on culture or the reference object, for example, then the scale works differently and no common frame of reference exists. In this case, measures are meaningful within the particular culture but not across cultures, for there is no cross-cultural equivalence of the instrument. In other words, the items are emic (Berry 1980) and their meaning depends on the particular cultural framework. Technically, this is called differential item functioning (DIF). If a set of items proves to be invariant, however, this set can be used to define the latent dimension across groups and items exhibiting DIF can be linked to this common scale based on group-specific item parameters. This is very similar to the idea of partial invariance in multigroup structural equation modeling (see Steenkamp and Baumgartner 1998). Consequently, non-etic items can be considered as well, provided etic items exist. Even fully emic items, that is, items that are only employed in one group, can be linked to the common scale (see Table 1, pp. 23–25). This can be done very easily using the Rasch model because missing data represents no problem in this respect.

The Rasch Model and Existing Scales

The analysis of existing scales often raises the problem of a limited variation in the items' locations as a consequence of their "classical" genesis. Nevertheless, in our view, it is important to reanalyze existing scales to reveal potential hidden deficiencies. Following the Rasch approach, the specification of the domain of the construct as advocated by Churchill (1979) should not only deal with the facets that belong to the construct in mind, but also specify the meaning of different levels on the continuum. The items of the original pool should reflect different degrees of the construct. Fit of the data to the model then establishes construct validity. In contrast to the traditional understanding of measurement, content validity and construct validity are much more interrelated because it is the item content that determines the item location. Is there any point, then, in reexamining scales? Should these scales not be evaluated according to classical standards? In our opinion, an attentive reanalysis is crucial because the standard way of how these scales are used—data collected, items scores added up, scores taken as measures—actually means that we presume the scales to be Rasch scales. In the ensuing study, we will therefore apply the Rasch model to investigate the measurement properties of an established advertising response scale (Schlinger 1979).

The Rasch Model for Polytomous Data

Since the response format of the current scales is polytomous, the dichotomous model cannot be applied. We decided against dichotomizing the polytomous data because of the potential loss of information, as well as to avoid theoretical problems (Andrich 1995a, 1995b). We therefore use the general Rasch model for polytomous data (Andrich 1978a, 1978b, 1988; Masters 1982) (see Figure 1). This model provides additional threshold parameters characterizing the transition points between two adjacent response categories. A polytomous item with *m* categories requires *m* – 1 threshold parameters. The average of the threshold parameters can be regarded as an overall location of the item. Since the threshold parameters of an item are estimated independently of each other in the Rasch model, they can be in any order. Thresholds that do not reflect the assumed order are called reversed thresholds. Reversed thresholds should not be ignored because they hint at problems underlying the process of responding to the item (Andrich 1995a). In the context of attitude measurement, reversed thresholds typically occur when more categories are provided than respondents actually use.

In such cases, the categories affected should be collapsed, as a different scoring of the categories is not justified when thresholds are reversed (Andrich 1995a). Ideally, new data should be collected on the basis of the adapted number of distinct categories. In practice, this may lead to a different number of categories for each item. Since this will be confusing for the respondent, post hoc rescoring appears to be the better choice in most cases.

Comparing Rasch Analysis with Multigroup CFA

Both approaches are concerned with the assessment of a common measurement model across different groups (i.e., populations). First, the factor analytic method investigates the invariance of factor loadings. Factor loadings are allowed to vary from item to item but are set equal across populations. The Rasch approach postulates equal discrimination across all items to keep the item-characteristic curves nonintersecting. Hence, the Rasch model is more demanding in this respect and there is no way to allow for different discrimination in different populations. Second, the factor analytic approach examines additive biases by including intercept parameters in the measurement equations. The concept of partial invariance allows for some items having unequal estimates of factor loadings across groups. In the Rasch analysis, this idea is paralleled by the possibility of splitting items functioning differently across groups as indicated by a DIF analysis.

FIGURE 1
General Polytomous Rasch Model

$$P(a_{vi} = x \mid \beta_v, \tau_{ij}\, j = 1\ \ldots\ m, 0 < x \leq m) = \frac{e^{\left(\sum_{j=1}^{x} - \tau_{ij}\right) + x \cdot (\beta_v - \delta_i)}}{\Upsilon}$$

with: $$\Upsilon = 1 + \sum_{k=1}^{m} e^{\left(\sum_{j=1}^{k} - \tau_{ij}\right) + k \cdot (\beta_v - \delta_i)}$$

β_v: person *v* location parameter
δ_i: item *i* location parameter
τ_{ij}: threshold *j* of item *i* parameter
m: maximum score, that is, number of categories – 1
a_{vi}: answer of person *v* to item *i* (item score)

Source: Andrich 1988, p. 366.

A theoretical problem lies in the relation of the intercept values and the factor loadings. Since the item scores are bounded between, for example, 1 and 5, the relation of the latent variable and the item score becomes increasingly nonlinear when approaching the extreme scores.

While both approaches serve the same purpose, there are several important differences that make Rasch analysis an attractive alternative. First, it is consistent with an axiomatic framework of measurement. This ensures that we achieve a level of measurement that is, in principle, comparable to measurement in the natural sciences. Second, it does not treat the manifest item scores as being linearly related to the measure but considers them to be nominally different responses with a hypothesized order. Third, the model estimates item and person parameters that best explain the manifest responses. If the measurement instrument works properly, the estimation of item parameters does not depend on the specific sample used and unbiased estimates of item properties may be obtained from unrepresentative samples (Embretson and Reise 2000, p. 23). Fourth, tests of fit investigate whether the data fit the model sufficiently to estimate measures. Fifth, the range of different item locations lends meaning to different levels on the dimension of interest and thereby enhances the interpretation of person measures. Table 1 presents a detailed comparison of Rasch and CFA approaches.

Integrating Rasch Analysis and Classical Test Theory

Our exposition of the Rasch model vis-à-vis the standard approach and Singh's (2004) comparison of classical test theory (CTT) and item response theory (IRT) might suggest that

TABLE 1
Confirmatory Factor Analysis (CFA) Versus Rasch Approaches: Differences and Measurement Implications

	CFA	Rasch
1. Fundamental and theoretical issues of measurement		
Concept of measurement	• Based on CTT. • Numbers are assigned to respondents attributes (Stevens 1946, 1951).	• The measure of a magnitude of a quantitative attribute is its ratio to the unit of measurement, the unit of measurement is that magnitude of the attribute whose measure is 1 (Michell 1999, p. 13). • Measurement is the process of discovering rations rather than assigning numbers. • Rasch model is in line with axiomatic framework of measurement. • Principle of specific objectivity.
Model	$\chi_\mathrm{i} = \tau_i + \lambda_{ij}\xi_\mathrm{j} + \delta_i$ χ_i . . . manifest item score τ_i . . . item intercept parameter λ_{ij}. . . factor loading of item i at factor *j* ξ_j . . . factor score of factor *j* δ_i . . . stochastic error term	For dichotomous data: $P(\alpha_{vi} = 1) = e^{(\beta v - \delta i)} / [1 + e^{(\beta v - \delta i)}]$ α_{vi} . . . response of person *v* to item i β_v . . . person location parameter δ_i . . . item location parameter (endorsability)
Relationship of measure and indicators (items)	• Measure is directly and linearly related to the indicators. • Hence, the weighted raw score is considered to be a linear measure.	• Probability of a response is modeled as a logistic function of two measures, the person parameter βv and the item location (endorsability) δi. • Raw score is not considered to be a linear measure, transformation of raw scores into logits (Wright 1996, p. 10).
In/dependence of samples and parameters	Parameters are sample-dependent, representative samples are important.	Item parameters are independent of sample used (subject to model fit and sufficient targeting).
2. Item selection and sampling (scale efficiency) issues		
Item selection	• Items selected to maximize reliability, leads to items that are equivalent in terms of endorsability, which plays no explicit role in CTT. • Favors items that are similar to each other (see bandwidth-fidelity problem; Singh 2004).	• Items are selected to cover a wide range of the dimension (see "bandwidth"; Singh 2004). • Endorsability of item plays a key role.
Item discrimination	• Discrimination varies from item to item, but is considered fixed within an item.	• Discrimination is equal for all items to retain a common order of all items in terms of endorsability for all respondents. • Discrimination varies within an item (concept of information that equals $P[\alpha_{vi} = 1]$ * $\mathrm{P}[\alpha_{vi} = 0]$ in the dichotomous case), it reaches its maximum at $\beta_v = \delta_i$.

(*continues*)

TABLE 1 (*continued*)

	CFA	Rasch
Targeting	Items that are off-target may even increase reliability and feign a small standard error that can actually be quite large.	Items that are off-target provide less information, standard errors will increase, and the power of the test of fit will decrease.
Standard error of measurement	Based on reliability, assumed to be equal across the whole range.	Based on the information, the items yield for a specific person.
Sample size	The required sample size mirrors recommendations for SEM. SEM is not appropriate for sample sizes below 100. As a rule of thumb, sample sizes of greater than 200 are suggested (Boomsma 1982; Marsh, Balla, and McDonald 1988). Bentler and Chou (1987) recommend a minimum ratio of 5:1 between sample size and the number of free parameter to be estimated.	In general, the sample sizes used in SEM are sufficient, but insufficient targeting increases the sample size needed. According to Linacre (1994), the minimum sample size ranges from 108 to 243, depending on the targeting with *n* = 150 sufficient for most purposes (for item calibrations stable within +/–.5 logits and .99 confidence).
Distribution of persons	Commonly assumed to be normal.	Irrelevant due to specific objectivity (subject to sufficient targeting).
Missing data	Problematic, missing data has to be imputed; deleting persons may alter the standardizing sample, and deleting items may alter the construct. Pairwise deletion biases the factors (Wright 1996, p. 10).	Estimation of person and item parameters not affected by missing data (except for larger standard errors).
Interpretation of person measures	Usually in reference to sample mean.	In reference to the items defining the latent dimension.
3. Dimensionality issues		
Multidimensionality	Multidimensionality easily accounted for.	A priori multidimensional constructs are split up into separate dimensions.
Directional factors	Sensitivity to directional factors (Singh 2004) in case of items worded in different directions.	Low sensitivity to directional factors (Singh 2004).
4. Investigation of comparability of measures across groups		
Assessment of scale equivalence	• Multigroup analysis. • Equivalence statements of parameters estimated across groups.	• DIF capitalizing on the principle of specific objectivity. • Analysis of residuals in different groups.
Incomplete equivalence	Partial invariance (for group-specific items, separate loadings and/or intercepts are estimated).	Item split due to DIF (for group-specific items, separate item locations are estimated).
Typical sequence and principal steps of analysis	• Estimation of baseline model (group-specific estimates of loadings and item intercepts).	• Estimation of model across groups. • Collapsing of categories if necessary.

	• Equality constraints imposed on loadings (metric invariance). • Equality constraints imposed on intercepts (scalar invariance). • Selected constraints lifted if necessary (partial invariance).	• Assessment of fit. • Assessment of DIF. • Items displaying DIF are split up if necessary.
Etic versus emic	• In principle, etic-oriented approach. A common set of invariant items is indispensable. • Concept of partial invariance allows for equal items functioning differently. • Emic items, that is, items confined to one group, can be considered but technical set-up complicated compared with Rasch analysis.	• In principle, etic-oriented approach. A common set of invariant items is indispensable. • Accounting for DIF by splitting the item allows for equal items functioning differently. • Emic items, that is, items confined to one group, can be considered very easily because handling of missing data is unproblematic compared with CFA.

Notes: CTT = classical test theory; SEM = structural equation modeling; DIF = differential item functioning.

these approaches are fundamentally irreconcilable. Despite significant theoretical differences, however, this is not necessarily the case. Rasch analysis begins with the assessment of the fit of the data to the model by examining whether the responses match the expected patterns and whether the hypothesis of an underlying latent variable, which is quantitative, is viable. As previously discussed, this is tantamount to the scientific task of measurement. Once this has been shown, the procedure of deriving actual measures is quite similar to the standard CTT approach. The item scores are added up to a total score. Following the standard approach, it is common to weight the item scores according to their factor loadings. In factor analysis, this is achieved by estimating factor score regression coefficients. It should be noted, however, that in its most stringent form, the CFA approach assumes parallel items, which are then added up equally weighted. In Rasch modeling, the items are also added up weighted equally since they are assumed to be parallel subject to item fit. If order is all we need, the total score is sufficient. However, since the metric of the raw score is not linear, it has to be converted to a linear measure as soon as metric interpretations are intended. This also allows for comparisons of measures derived from different sets of items.

Consequently, rather than arguing that the Rasch model provides an alternative to the standard model of testing, we contend that the Rasch model is a useful supplement to CTT, because it extends the analytic process of measurement *before* item scores are condensed to scores (investigation of response patterns) and *afterward* (transformation of the nonlinear raw score into a linear measure). Hence, we claim that CTT can be seen as an abridged version of Rasch modeling.

Given the widespread use of structural equation modeling and CFA, why should researchers be tempted to use Rasch? The reasons are twofold: theoretical and practical. From a methodological point of view, the theoretical foundations of Rasch seem to be superior to those of CTT. While CTT presumes quantity, Rasch investigates the consequences of quantity, namely, specific patterns in the responses. If these patterns are absent, Rasch suggests refraining from inferring quantities in the data, even if the variables are reasonably correlated. From a practical point of view, this is clearly very advantageous. It is important to bear in mind, however, that the process is still essentially guided by theory and does not rely solely on researcher intuition. A direct consequence of specific objectivity is the sample independence of item parameters. Thus, representative samples are not needed with Rasch. Given timing and costing issues involved in the identification of representative samples, this is clearly an advantage, particularly when dealing with scale development projects at the item-generation and testing stages. Similarly, if testing for a scale's suitability for a new context, preliminary testing can be undertaken without the need to build on representative samples. Another major practical advantage is related to the testing of comparability of measures. Using DIF, equivalence issues and comparability of measures can be investigated quickly and efficiently. Within the RUMM (Rasch Unidimensional Measurement Models) software, for example, accounting for DIF works very easily by splitting items affected by DIF without the need to set up new data sets. Finally, Rasch measurement assumes responses on an ordinal level, thus avoiding the virtual discussion of Likert scales as "quasi-metric" scales, or the usual problems associated with non-normality of the data.

ANALYSIS

We use RUMM 2020 software in our analysis (Andrich, Sheridan, and Luo 2003b), which allows for the estimation of dichotomous as well as polytomous models among models for other purposes, such as multiple-choice data (see Appendix E for an example of the software). The program employs a conditional pairwise estimation approach that allows for item-parameter estimation while conditioning out the person parameters (see Andrich, Sheridan, and Luo 2003a for further details). For dichotomous items, the pairwise approach has been shown algebraically to provide estimates consistent with those yielded by conditional maximum likelihood estimation (Zwinderman 1995). For the polytomous case, the simulation study by Andrich, Sheridan, and Luo (2003a) provides evidence of the consistency of parameter estimates.

Empirical Investigation of the Measurement Properties

The empirical data analysis begins with the pooled data set. If it turns out that equivalence may not be achieved by accounting for unique item-parameter estimates for different groups, data sets are split up and analyses are carried out within subsamples. The path of our analyses, carried out separately for each dimension, takes the following sequence:

1. Ensuring That the Item Response Scale Works as Intended

First, the order of the thresholds is investigated. If disordered thresholds occur, the scoring of the item responses is not justified. Figure 2 depicts the category characteristic curves for an item measuring confusion (the scoring always starts with 0). The middle category (scored 2) never becomes the most likely option, and the transition point of switching from 1 to 2 (1.07) is further up the scale than the transition point between 2 and 3 (–.56). In such a case, adjacent categories are collapsed until a proper order of the thresholds is established. In other words, two categories are scored identically.

2. Testing the Item Fit to the Model

After accounting for reversed thresholds, misfit is assessed by a test of fit comparing the expected item score (based on the probabilities implied by the model parameter estimates) and the actual scores. RUMM2020 provides two sorts of statistical tests of the fit statistic. The χ^2 test is constructed as an approximate χ^2 (Andrich, Sheridan, and Luo 2003a). It assesses the difference between expected proportion, that is, the probability implied by the model, and the actual proportion at the group level, that is, a group of respondents with similar but not necessarily identical locations. A newly introduced *F* statistic accounts for individual differences. It is, therefore, the more sensitive fit statistic. For this reason, we base our decisions in terms of item fit on the *F* statistic.

Nevertheless, fit statistics should not be judged absolutely. If the probability of an item's *F* value is slightly below the α acceptance level (e.g., .008 in the case of an α of .01), it is unreasonable to discard the item simply because of the test of fit without further consideration. Content, for instance, should always be taken into consideration. If the item's content is unique and no other item has a similar location, discarding the item would lead to a gap in the scale that might be problematic. Another point is related to other statistical evidence. If the discrimination of the item is acceptable (i.e., the fit residual provided by RUMM lies within -2 and $+2$) and the next item in terms of the fit statistic shows a considerably worse misfit (e.g., p = .00001), then the item should be retained. This is similar to the scree-plot in factor analysis where the meaningfulness of a factor is not assessed only by an absolute interpretation of its eigenvalue, but by the eigenvalue relative to the eigenvalue of the factor extracted next.

Figure 3 shows the plot of the expected value depending on the person location on the latent dimension. The actual mean scores of six groups of respondents are represented by the dots. In this case, the dots are very close to what is expected according to the model. Consequently, the fit statistics ("ChiSq[Pr]" and "F[Pr]") are nonsignificant.

Actual discrimination can be examined by an additional fit statistic, a residual test-of-fit ("FitRes") reported by RUMM 2020. Andrich, Sheridan, and Luo state, "A 'very positive' value implies poor discrimination; a 'very negative' value implies too good a discrimination" (2003a, p. 25). The statistic is constructed as a standard normalized residual having an expected mean of 0 and a standard deviation of 1, but is not perfectly normally distributed. This statistic is very similar to the outfit-statistic provided by, for example, Winsteps (Linacre 1994). One should be concerned with values <-2 or, in particular, $>+2$.

Item misfit normally entails the deletion of the item. However, sometimes there are particular subpopulations causing the item to malfunction while the item works properly for other subpopulations. Alternatively, the item may function differently, that is, it can be endorsable more easily in one population compared with another. This question is addressed in step 3.

FIGURE 2
Category Characteristic Curves

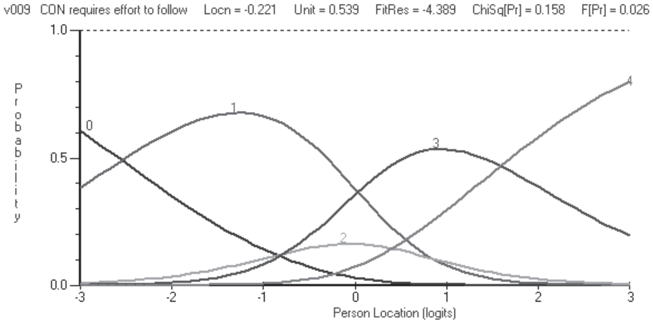
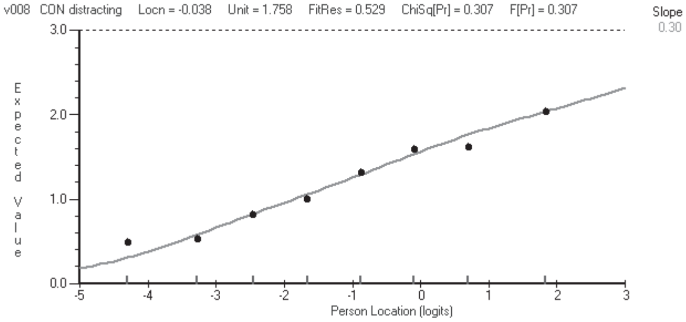


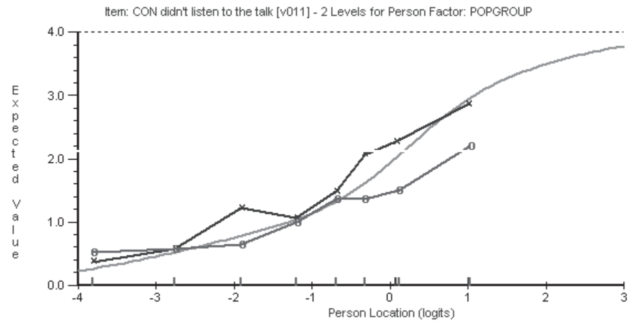
FIGURE 3
Item Characteristic Curves



3. Assessing Differential Item Functioning (DIF)

Assessing DIF actually serves two purposes. First, it is a test-of-fit because the model entails parameter estimates that should be independent of the respondents. This hypothesis is tested by a DIF analysis. Second, the DIF test explicitly examines the instruments equivalence for predefined groups of respondents. It is then up to the researcher to account for DIF by estimating item parameters for each group separately. In RUMM 2020, DIF is assessed by a two-way analysis of variance of the residuals (i.e., the difference of expected scores and actual scores) implemented in RUMM 2020 (Andrich, Sheridan, and Luo 2003a, 31f.). For this purpose, respondents are grouped into classes of ±60 persons each along a continuum with the class intervals being the first factor. The second factor is represented by the groups for which DIF is to be investigated, such as the country the respondents come from, or their gender. Significance of the first factor mirrors the general fit of the item, whereas significance of the second factor implies DIF. The reason for the latter can be illustrated as follows. Assume the true item location is different for two

FIGURE 4
Item Characteristic Curve, Analysis of Differential Item Functioning (DIF)



```
  SOURCE                S.S      M.S     DF      F-RATIO     Prob
-----------------------------------------------------------------------
BETWEEN                64.078    4.272   15
  ANOVA-Fit[CInt]      19.103    2.729    7     3.556773   0.000925
  TOTAL Item DIF       44.975    5.622    8     7.327109   0.000000
    DIF[popgroup]      19.716   19.716    1    25.696800   0.000007
    popgroup-by-CInt   25.259    3.608    7     4.702868   0.000031
```

Note: ANOVA = analysis of variance.

groups. If the location is estimated with pooled data, then the estimate lies in between the true locations for each group. Consequently, the mean of the residuals would then be positive in one group and negative in the other. Since severe DIF can also lead to item misfit, steps 2 and 3 are closely related.

Figure 4 illustrates the issue of DIF with an empirical example of a confusion item. The two lines connecting the mean scores of the two population groups clearly differ. The analysis of variance shows that the difference is highly significant.

4. Discarding Items, Accounting for DIF, or Separate Analyses

If necessary, misfitting items are deleted. In the case of DIF, the item affected can be split up and parameters will be estimated that pertain to specific subpopulations. This approach is only feasible, however, if there are enough items free of DIF to establish a common scale across subpopulations. Otherwise, separate analyses confined to part of the samples involved might be appropriate.

An Empirical Example: Applying the Rasch Model to Schlinger's Confusion Scale

By way of illustration, we use commercial data collected by a leading research company in South Africa. This analysis allows us to test for DIF in terms of different advertisements (four financial services commercials have been used, but in each instance, exactly the same advertisements were used for both population groups, with only the voice-overs being changed) and in terms of differences between nonindigenous (NI) and indigenous (I) South Africans (100 respondents per population group per advertisement). For a description of the commercials used, refer to Appendix A. Thus, our research design responds to the recent call of Lenartowicz, Johnson, and White (2003) for greater attention to be paid to intracountry cultural variation in international research. In addition, the invariance property can be tested for age (grouped into four classes), gender, and the degree to which respondents like the commercial. The liking of the commercial is assessed on a 10-point scale, which we have subsequently dichotomized to "like" versus "dislike." Accordingly, advertisement 1 is disliked by both groups, advertisement 4 is liked by both groups, advertisement 2 is disliked by nonindigenous but liked by indigenous viewers, and advertisement 3 is disliked by indigenous but liked by nonindigenous viewers. By using all four standardized advertisements from the same product category but with differing likeability scores, we are attempting to control for object (advertisement) variation.

The fit of the scale intended to measure confusion consisting of four items shows insufficient fit (χ^2 = 63.70, *df* = 28, *p* = .001) and reversed thresholds throughout all items (see Appendix B). After collapsing adjacent categories, the items do not show reversed thresholds any longer. However, the fit does not improve (χ^2 = 73.75, *df* = 28, *p* < .0001) (see Appendix C). A DIF analysis of the misfitting items, that is, 10 and 11, based on population group and commercial shows that no item displays DIF based on the commercial, but item 11 exhibits population group DIF. After splitting item 11, the item fits for nonindigenous viewers, but not for indigenous respondents. In addition, the misfit of item 10 does not improve. On the contrary, item 9 now misfits and shows population group DIF. However, after accounting for DIF for item 9 as well, both item 9 and 11 do not fit for indigenous respondents. A now significant DIF analysis based on the commercial could mean that the different commercials make the scale work differently. Item 10 still misfits but displays no DIF. In summary, this suggests serious problems with the confusion scale. We now proceed to separate nonindigenous from indigenous respondents. For nonindigenous viewers, a common model covering all four commercials seems tenable (χ^2 = 45.45, *df* = 28, *p* = .02, person-separation index .87). Only item 9 ("requires a lot of effort to follow") lacks satisfactory fit due to some overdiscrimination (see Tables 2 and 3).

For indigenous viewers, item 11 remains problematic, displaying misfit and DIF based on the commercial. However, neither splitting the item according to commercial nor a deletion of the item leads to a better model fit. Consequently, no final model for indigenous viewers covering all four commercials can be established. Separate models for each commercial fit reasonably well, but sample sizes are a little small to yield a high power of the test of fit.

Although the patterns of item location estimates are vastly different between commercials, there are some similarities. For example, item 9 ("requires a lot of effort") tends to be the

TABLE 2
Item Parameter Estimates for the Confusion Dimension, Nonindigenous (NI) Population Only

Item	Wording (abbreviation)	Location	*SE*	FitResid	*F* statistic	Probability
8	Distracting	–.038	.105	.529	1.180	.31
9	Required a lot of effort to follow	.395	.105	–3.289	2.946	.002
10	Too complex	.290	.105	–1.807	2.595	.01
11	Busy watching the screen, didn't listen to the words	–.646	.097	.875	1.392	.19

Notes: Viewer response profile, confusion, NI only; $\chi^2 = 45.45$, $df = 28$, $p = .02$.

TABLE 3
Item Parameter Estimates for the Confusion Dimension, Indigenous Respondents Only

Model		Commercial 1	Commercial 3	Commercial 2	Commercial 4
Indigenous respondents only	Model fit	$\chi^2 = 7.07$ $df = 8$ $p = .53$	$\chi^2 = 11.70$ $df = 8$ $p = .17$	$\chi^2 = 8.42$ $df = 8$ $p = .39$	$\chi^2 = 6.39$ $df = 8$ $p = .60$
	Sample size	91	87	87	96
	Person separation index	.73	.85	.82	.73
Item location estimates					
8	Distracting	.557	–.240	–.076	.753
9	Required a lot of effort to follow	–1.970	–.347	–.712	–.834
10	Too complex	.897	.928	.277	1.109
11	Busy watching the screen, didn't listen to the words	.516	–.341	.511	–1.029

Notes: Viewer response profile, confusion, indigenous respondents only; comparison of models.

item most agreed with, whereas item 10 ("too complex") is problematic. It is highly plausible that "requires a lot of effort" is easier to agree with than "too complex" because "requires a lot of effort" implies "complex," whereas "too complex" expresses even more complexity. Surprisingly, no such relation can be found for nonindigenous respondents; for them, both items are about equally endorsable. In summary, this raises serious issues in terms of the construct validity of the confusion scale. Nonindigenous viewers seem to interpret the items rather similarly, that is, they do not represent a clear hierarchy from less confusion to more confusion. In contrast, for indigenous respondents, there basically is such a hierarchy, but it depends on the specific commercial.

Finally, CFA and multigroup invariance testing was conducted to assess the structure of the confusion dimension for nonindigenous (NI) and indigenous (I) South Africans. Despite good model fit in the separate CFAs, invariance testing demonstrated that the items were not fully metrically invariant (Steenkamp and Baumgartner 1998). Partial equivalence could be established, however. The results (see Appendix D) are in line with findings from the Rasch analysis in that they support differential item functioning.

CONCLUSIONS

Our analysis of Schlinger's confusion scale uncovers several psychometric problems. For indigenous respondents, the items appear to be highly stimulus idiosyncratic; in other words, their functioning depends heavily on the particular commercial being evaluated. For nonindigenous respondents, however, the scale is stable across commercials. Some degree of DIF occurs, though, but can be accounted for. Yet it should be noted that accounting for DIF, elegant and straightforward as it may seem, bears some theoretical problems. It can be seen as "a threat to the desirable invariance of the measurement instrument across persons" (Molenaar 1997, p. 40). Whenever a person's measure depends on the person's group membership, this is inevitably associated with a "relaxed" understanding of specific objectivity. In addition, it incurs the risk of chance capitalization and unstable parameter estimation (Molenaar 1997). Following the traditional approach of scale development this may seem irrelevant, for inter-item correlation is essentially sufficient. From a Rasch perspective, however, construct validity depends on a theory-driven hierarchy of items that make up a latent dimension.

As a general conclusion, we suggest rethinking construct and cross-cultural validity. Our finding that the meaning of some items depends largely on the particular advertisement demonstrates that different populations of respondents are not the *only* source of variance of a measurement scale. Do measures of a construct (e.g., entertainment) based on different *objects* (e.g., advertisements) bear the same meaning? This needs to be explicitly raised and empirically tested. In his C-OAR-SE approach to scale development, Rossiter (2002) also emphasizes the importance of the object. He even makes the object a part of the definition of the construct. If one defines the object too narrowly, however, that is, as one specific advertisement, then any comparison of measures across different objects is ruled out by definition, because the measures are related to different constructs, even when they are provided by identical items. In contrast, we suggest making it a matter of empirical investigation how far the construct's frame of reference actually reaches in terms of objects. The Rasch modeling approach allows one to test for comparability across objects (i.e., commercials) as easily as it facilitates the assessment of cross-cultural equivalence. Both issues can be the subject of a differential item functioning analysis, provided the data set comprises more than one different object.

Finally, one has to question the value of such sophisticated psychometric analyses to both advertising science and practice. For what it is worth, we believe that scientific research should always consider the latest methodology and, where appropriate, employ the most advanced models available. In this regard, true score theory is certainly not the most advanced way to tackle measurement problems. Admittedly, measurement in the social sciences is a difficult undertaking. However, this fact alone should not keep us from investigating measurement instruments thoroughly. Just because measurement is difficult does not mean that we have a license to handle it in a haphazard fashion.

The Rasch model provides a sound theoretical basis for measurement. A goal of this paper has been to overcome some of the obstacles to its more mainstream use. Once the analyst has the software, the actual mechanics of Rasch analysis is certainly no more complicated than any structural equation modeling problem. In contrast to more complex IRT models, Rasch analysis does not even require large sample sizes. The sample sizes usually deemed appropriate for CFA are perfectly suitable for Rasch model estimation in most cases. Granted, the Rasch model represents a stringent and demanding test for any scale. In our view, however, it is better to know where the problems and the limitations of a scale lie and then use the scale accordingly than to assume that the scale works well, when in fact, the purported quantification is doubtful, to say nothing about the consequences for substantive conclusions based on suspect measurement.

We also have to keep in mind that scale development is not an academic pastime. We take the responsibility for properties of our published scales when practitioners make use of them. This leads us to the consequences of scale development and analysis for marketing practice. Practitioners will probably not carry out complicated psychometric analyses. After all, that is the responsibility of scientific research. Practitioners need to rely on the quality of a scale, and they need to be aware of the limitations of a scale. For example, the dependency of some dimensions on the particular commercial suggests that deciding between different commercials should never be made solely on the basis of the scale. On the other hand, cross-cultural comparisons are justified in many cases, and can be made reliably. Rasch analysis does not force the practitioner to use the linear measures, that is, the person location estimates, although simple translation schemes from raw score to measure can easily be provided. If ordinal data are sufficient (i.e., when commercials are to be ranked), the nonlinear raw score is perfectly suitable for inferences. Over a relatively wide range, the raw score is almost equal interval scaled anyway. It should be noted that these properties of the raw score, that is, the meaningfulness of its unweighted composition and the near interval scale around the center of the instrument, are a consequence of fit to the Rasch model. In other words, to confine oneself to the raw score does not release the scientist from scale analysis.

To conclude, our analyses reveal that the confusion scale has some problems. We also show where further improvements can be made, however. A quote from Molenaar provides a lucid summary of our conclusions: "One thing must be clear: the high correlations and robust conclusions . . . are by no means a good reason for sloppy modeling or sloppy measurement. Careful modeling and careful measurement bear some resemblance to airbags and safety belts in a car: they are somewhat costly, and one rarely needs them, but when one does, their presence matters an awful lot" (1997, p. 492).

NOTE

1. The interested reader is referred to Michell (1990) and Karabatsos (2001) for an in-depth discussion of the cancellation conditions.

REFERENCES

Aaker, David A., and Douglas M. Stayman (1990), "Measuring Audience Perceptions of Commercials and Relating Them to Ad Impact," *Journal of Advertising Research,* 30 (4), 7–17.

Andrews, J. Craig, Srinivas Durvasula, and Richard G. Netemeyer (1994), "Testing the Cross-National Applicability of U.S. and Russian Advertising Belief and Attitude Measures," *Journal of Advertising,* 23 (1), 71–82.

Andrich, David (1978a), "Application of a Psychometric Rating Model to Ordered Categories Which Are Scored with Successive Integers," *Applied Psychological Measurement,* 2 (4), 581–594.

——— (1978b), "A Rating Formulation for Ordered Response Categories," *Psychometrika,* 43 (4), 561–573.

——— (1988), "A General Form of Rasch's Extended Logistic Model for Partial Credit Scoring," *Applied Measurement in Education,* 1 (4), 363–378.

——— (1995a), "Further Remarks on Non-Dichotomization of Graded Responses," *Psychometrika,* 60 (1), 37–46.

——— (1995b), "Models for Measurement, Precision and the Non-Dichotomization of Graded Responses," *Psychometrika,* 60 (1), 7–26.

———, Barry S. Sheridan, and Guanzhong Luo (2003a), "Displaying the Rumm2020 Analysis," in RUMM Laboratory working paper, Perth, Western Australia.

———, ———, ——— (2003b), "Rumm2020: Rasch Unidimensional Measurement Models," in RUMM Laboratory working paper, Perth, Western Australia.

Bagozzi, Richard P., and Youjaee Yi (1988), "On the Evaluation of Structural Equation Models," *Journal of the Academy of Marketing Science,* 16 (1), 74–94.

Balasubramanian, Siva K., and Wagner A. Kamakura (1989), "Measuring Consumer Attitudes Toward the Marketplace with Tailored Interviews,"*Journal of Marketing Research,* 26 (3), 311–326.

Balnaves, Mark, and Peter Caputi (2001), *Introduction to Quantitative Research Methods: An Investigative Approach,* London: Sage.

Bentler, Peter M., and Chih-ping Chou (1987), "Practical Issues in Structural Modeling," *Sociological Methods and Research,* 16 (1), 78–117.

———, and Eric J. C. Wu (2002), *Eqs 6 for Windows User's Guide,* Encino, CA: Multivariate Software.

Berry, John W. (1980), "Introduction to Methodology," in *Handbook of Cross-Cultural Psychology,* vol. 2, *Methodology,* Harry C. Triandis and John Berry, eds., Boston: Allyn and Bacon, 1–28.

Birnbaum, Allan (1968), "Some Latent Trait Models and Their Use in Inferring an Examinee's Ability," in *Statistical Theories of Mental Test Scores,* Frederic M. Lord and Melvin R. Novick, eds., Reading, MA: Addison-Wesley, chapters 17–20.

Bond, Trevor G., and Christine M. Fox (2001), *Applying the Rasch Model: Fundamental Measurement in the Human Sciences,* Mahwah, NJ: Lawrence Erlbaum.

Boomsma, Anne (1982), "The Robustness of Lisrel Against Small Sample Sizes in Factor Analysis Models," in *Systems Under Indirect Observation: Causality, Structure, and Prediction* (part 1), Karl G. Jöreskog and Herman Wold, eds., Amsterdam: North-Holland.

Burke, Marian Chapman, and Julie A. Edell (1982), "Exploring the Multidimensional Nature of Attitude Toward the Ad," unpublished working paper, Duke University.

Christensen, Karl Bang, and Jakob Bue Bjorner (2003), "SAS Macros for Rasch-Based Latent Variable Modeling," Technical Report no. 03/13, University of Copenhagen, Department of Biostatistics, available at www.biostat.ku.dk/publ-e.htm (accessed September, 13, 2004).

Churchill, Gilbert A. (1979), "A Paradigm for Developing Better Measures of Marketing Constructs," *Journal of Marketing Research,* 16 (1), 64–73.

de Mooij, Marieke K. (1994), *Advertising Worldwide: Concepts, Theories and Practice of International, Multinational and Global Advertising,* 2nd ed., New York: Prentice Hall.

Deshpande, Rohit, and Douglas M. Stayman (1994), "A Tale of Two Cities: Distinctive Theory and Advertising Effectiveness," *Journal of Marketing Research,* 31 (1), 57–64.

Durvasula, Srinivas, Craig J. Andrews, Steven Lysonski, and Richard G. Netemeyer (1993), "Assessing the Cross-National Applicability of Consumer Behavior Models: A Model of Attitude Toward Advertising in General,"*Journal of Consumer Research,* 19 (4), 626–636.

Embretson, Susan E., and Steven P. Reise (2000), *Item Response Theory for Psychologists,* Mahwah, NJ: Lawrence Erlbaum.

Ewing, Michael T., Albert Caruana, and Andy Teo (2002), "Towards the Development of a Scalar Equivalent Etic Multicultural Advertising Response Scale (Mars)," in *New Directions in International Advertising Research,* Charles R. Taylor, ed., Greenwich, CT: JAI Press, 1–8.

———, ———, and George M. Zinkhan (2002), "On the Cross-National Generalisability and Equivalence of Advertising Response Scales Developed in the USA," *International Journal of Advertising,* 21 (3), 323–344.

Guttman, Louis (1950), "The Basis for Scalogram Analysis," in *Measurement and Prediction: Studies in Social Psychology in World War II,* vol. 4, Samuel Andrew Stouffer, Louis Guttman, Edward A. Suchman, Paul F. Lazarsfeld, Shirley A. Star, and John A. Clausen, eds., Princeton: Princeton University Press, 60–90.

Hofstede, Geert (1984), *Culture's Consequences: International Differences in Work-Related Values,* Cross-Cultural Research and Methodology Series, Newbury Park, CA.: Sage.

James, David (2001), "Local Coke," *Business Review Weekly,* 23, 70–74.

Karabatsos, George (2001), "The Rasch Model, Additive Conjoint Measurement, and New Models of Probabilistic Measurement Theory,"*Journal of Applied Measurement,* 2 (4), 389–423.

Knight, Gary A., Richard A. Spreng, and Attila Yaprak (2003), "Cross-National Development and Validation of an International Business Measurement Scale: The Coiscale," *International Business Review,* 12 (5), 581–599.

Lenartowicz, Tomasz, James P. Johnson, and Carolyn T. White (2003), "The Neglect of Intracountry Cultural Variation in International Management Research," *Journal of Business Research,* 56 (12), 999–1008.

Linacre, John Michael (1994), "Sample Size and Item Calibration Stability," *Rasch Measurement Transactions,* 4, 328.

——— (1991–2004), *Winsteps: Rasch Model Computer Program,* Chicago: MESA.

Lutz, Richard J. (1985), "Affective and Cognitive Antecedents of Attitude Toward the Ad: A Conceptual Framework," in *Psychological Processes and Advertising Effects: Theory, Research, and Applications,* Linda F. Alwitt and Andrew A. Mitchell, eds., Hillsdale, NJ: Lawrence Erlbaum, 45–63.

Marsh, Herbert W., John R. Balla, and Roderick P. McDonald (1988), "Goodness-of-Fit Indexes in Confirmatory Factor Analysis: The Effect of Sample Size," *Psychological Bulletin,* 103 (3), 391–410.

Masters, Geoffrey N. (1982), "A Rasch Model for Partial Credit Scoring," *Psychometrika,* 47 (2), 149–174.

Mbigi, Lovemore (2000), *In Search of the African Business Renaissance: An African Cultural Perspective,* Randburg, South Africa: Knowledge Resources.

Michell, Joel (1990), *An Introduction to the Logic of Psychological Measurement,* Hillsdale, NJ: Lawrence Erlbaum.

——— (1999), *Measurement in Psychology: A Critical History of a Methodological Concept,* Cambridge: Cambridge University Press.

Molenaar, Ivo W. (1997), "Lenient or Strict Application of IRT with an Eye on the Practical Consequences," in *Applications of Latent Trait and Latent Class Models in the Social Sciences,* Jürgen Rost and Rolf Langenheine, eds., Münster: Waxmann, 38–49.

Onkvisit, Sak, and John J. Shaw (1987), "Standardized International Advertising: A Review and Critical Evaluation of the Theoretical and Empirical Evidence," *Columbia Journal of World Business,* 22 (3), 43–55.

Rasch, Georg (1960), *Probabilistic Models for Some Intelligence and Attainment Tests,* Copenhagen: Danish Institute for Educational Research.

——— (1961, 20/6–30/7 1960), "On General Laws and the Meaning of Measurement in Psychology," Berkeley Symposium on Mathematical Statistics and Theory of Probability, vol. 4, Berkeley: University of California Press, 321–333.

——— (1977), "On Specific Objectivity: An Attempt at Formalizing the Request for Generality and Validity of Scientific Statements," *Danish Yearbook of Philosophy,* 14, 58–94.

Rossiter, John R. (2002), "The C-OAR-SE Procedure for Scale Development in Marketing," *International Journal of Research in Marketing,* 19 (4), 305–335.

Salzberger, Thomas, Rudolf R. Sinkovics, and Bodo B. Schlegelmilch (1999), "Data Equivalence in Cross-Cultural Research: A Comparison of Classical Test Theory and Latent Trait Theory Based Approaches," *Australasian Marketing Journal,* 7 (2), 23–38.

Samejima, F. (1997), "Graded Response Model," in *Handbook of Modern Item Response Theory,* Wim J. Van der Linden and Ronald K. Hambleton, eds., New York: Springer, 85–100.

Schaffer, Bryan S., and Christine M. Riordan (2003), "A Review of Cross-Cultural Methodologies for Organizational Research: A Best-Practices Approach," *Organizational Research Methods,* 6 (2), 169–215.

Schlinger, Mary Jane (1979), "A Profile of Responses to Commercials," *Journal of Advertising Research,* 19 (2), 37–46.

Singh, Jagdip (2004), "Tackling Measurement Problems with Item Response Theory: Principles, Characteristics, and Assessment, with an Illustrative Example," *Journal of Business Research,* 57 (2), 184–208.

———, Roy D. Howell, and Gary K. Rhoads (1990), "Adaptive Designs for Likert-Type Data: An Approach for Implementing Marketing Surveys," *Journal of Marketing Research,* 27 (3), 304–321.

Sinkovics, Rudolf R., Thomas Salzberger, and Hartmut H. Holzmüller (1998), "Assessing Measurement Equivalence in Cross-National Consumer Behavior Research: Principles, Relevance and Application Issues," in *New Developments and Approaches in Consumer Behaviour Research,* Ingo Balderjahn and Claudia Mennicken, eds., London: MacMillan, 269–288.

Soutar, Geoffrey N., Richard Bell, and Yvonne Wallis (1990), "Consumer Acquisition Patterns for Durable Goods: A Rasch Analysis," *Asia Pacific International Journal of Marketing,* 2 (1), 31–39.

———, and Steven P. Cornish-Ward (1997), "Ownership Patterns for Durable Goods and Financial Assets: A Rasch Analysis," *Applied Economics,* 29, 903–911.

———, and Maria M. Ryan (1999), "People's Leisure Activities: A Logistic Modelling Approach," Australia and New Zealand Marketing Academy Conference (ANZMAC), Sydney, Australia, November 29–December 2.

Steenkamp, Jan-Benedict, and Hans Baumgartner (1998), "Assessing Measurement Invariance in Cross-National Consumer Research," *Journal of Consumer Research,* 25 (1), 78–90.

Stevens, S. S. (1946), "On the Theory of Scales of Measurement," *Science,* 103 (2684), 667–680.

——— (1951), "Mathematics, Measurement, and Psychophysics," in *Handbook of Experimental Psychology,* S. S. Stevens, ed., New York: Wiley, 1–49.

Taylor, Charles R. (2003), "New Books in Review: Cross-Cultural Survey Methods," *Journal of Marketing Research,* 40 (2), 246–247.

Tesio, Luigi (2003), "Measuring Behaviours and Perceptions: Rasch Analysis as a Tool for Rehabilitation Research," *Journal of Rehabilitation Medicine,* 35, 105–115.

Traub, Ross E. (1994), *Reliability for the Social Sciences: Theory and Applications,* Sage Measurement Methods for the Social Sciences, Thousand Oaks, CA: Sage.

Wright, Benjamin D. (1996), "Comparing Rasch Measurement and Factor Analysis," *Structural Equation Modeling,* 3 (1), 3–24.

———, and Geoffrey N. Masters (1982), *Rating Scale Analysis,* Chicago: Mesa Press.

Zinkhan, George M., and Scott Burton (1989), "An Examination of Three Multidimensional Profiles for Assessing Consumer Reactions to Advertisements," *Journal of Advertising,* 18 (4), 6–13.

———, and Claes Fornell (1985), "A Test of Two Consumer Response Scales in Advertising," *Journal of Marketing Research,* 22 (4), 447–452.

Zwinderman, Aeilko H. (1995), "Pairwise Parameter Estimation in Rasch Models," *Applied Psychological Measurement,* 19 (4), 369–375.

APPENDIX A

Description of the Four Commercials

Ad 1

A humorous ad for a bank's homeloan/mortgage, featuring both indigenous and nonindigenous actors. Opening voice-over: *You are about to buy a house; now who do you talk to?* Cut to very pessimistic "advice" from parents, friends, hairdresser. Cut to the interior of the bank. Return to voice-over: *Open one door for the financial help and advice you need to buy a home. Talk to the real experts. ABC people you can talk to.*

Ad 2

While not created by J. Walter Thompson, this insurance company ad fits the JWT definition of a vivid metaphor. It focuses on "life" and uses stunning scenes from nature revolving around birth, life cycles, and so forth. It also has powerful music. Voice-over: *Life is the greatest gift of all . . . but every creature on planet earth gets only one life. At XYZ, we'd like to help you make the most of your life-every step of the way.*

Ad 3

Is the same insurance company as ad 2, but this execution focuses on optimism and hope (thereby tapping into the positive psyche of the post-apartheid "new South Africa"). Shows powerful images of war scenes from the 1940s (World War II), the 1960s (Vietnam) . . . and then cuts to Germans tearing down the Berlin Wall, Nelson Mandela's release from jail, Reagan shaking hands with Gorbachev, and so forth. Voice-over: *Like you, XYZ has hopes and dreams for the future. We'd like to help you turn your hopes into achievements and your dreams into realities. We'd like to help you make the most of your life—every step of the way.*

Ad 4

Is for a different insurance company from ads 2 and 3. It uses "talking" toddlers. A female (baby) fortune-teller is about to consult a male (baby) client. She quickly replaces her crystal ball with a (corporate) blue clay ball, which expands and morphs to demonstrate how the one policy from '12'3 can cover all needs, and so forth. The indigenous version using indigenous babies and a traditional medicine women (*sangoma*) rather than the Western-style gypsy fortune-teller.

APPENDIX B

Item Parameter Estimates for the Confusion Dimension

Item	Wording (abbreviation)	Location	*SE*	FitResid	*F* statistic	Probability
8*	Distracting	.013	.056	–1.525	1.364	.20
9*	Required a lot of effort to follow	–.221	.049	–4.389	2.126	.03
10*	Too complex	.388	.059	–4.530	5.759	<.0001
11*	Busy watching the screen, didn't listen to the words	–.180	.051	1.413	2.566	.01

Notes: Viewer response profile, confusion, complete data set; χ^2 = 63.70, *df* = 28, *p* = .001.

* Reversed thresholds.

APPENDIX C

Item Parameter Estimates for the Confusion Dimension, After Rescoring Item Responses

Item	Wording (abbreviation)	Location	*SE*	FitResid	*F* statistic	Probability
8	Distracting	.016	.074	–1.134	1.391	.19
9	Required a lot of effort to follow	–.239	.066	–5.099	2.183	.029
10	Too complex	.486	.074	–4.523	6.080	<.0001
11	Busy watching the screen, didn't listen to the words	–.263	.068	1.456	3.738	<.001

Notes: Viewer response profile, confusion, complete data set; χ^2 = 73.75, *df* = 28, p < .0001.

APPENDIX D

Multigroup Confirmatory Factor Analyses (CFA) for Confusion

CFA and multigroup invariance testing was conducted to assess the structure of the confusion dimension for nonindigenous (NI) and indigenous (I) South Africans, separately. An initial CFA on the full sample indicated a very good fit to the data (see Table D.1 below). Using the maximum likelihood method, all loadings between the four items and the single latent factor were free parameters to be estimated and the variance was fixed for the latent factor (conventional CFA).

TABLE D.1

CFA for the Confusion Dimension, Full Sample

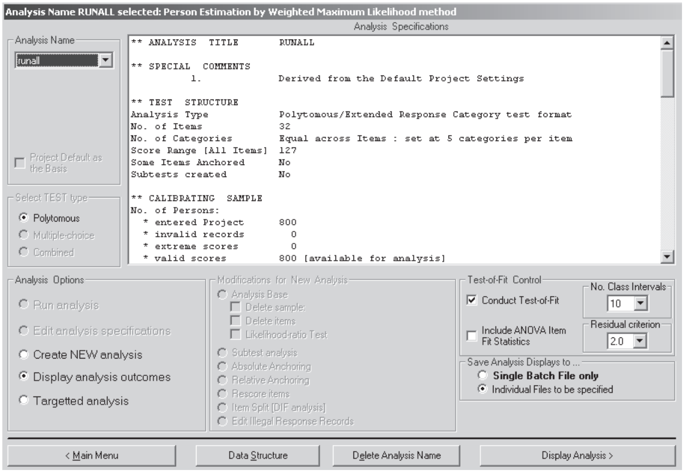
Id.	Item	μ	*SD*	$\lambda_{i,j}$	t value
conf1	It was distracting trying to watch the screen and listen to the words at the same time.	2.06	.90	.734	22.43
conf2	It required a lot of effort to follow the commercial.	2.12	1.05	.820	25.92
conf3	It was too complex. I was not sure what was going on.	1.89	.81	.806	25.35
conf4	I was so busy watching the screen, I didn't listen to the words.	2.09	.98	.606	17.55

Confusion dimension – α = .82, ρ = .83, AVE = .56

Notes: α = Cronbach's alpha; ρ = Jöreskog's rho measure of construct reliability; AVE = average variance extracted; μ = mean; *SD* = standard deviation; λ = standardized path coefficient; χ^2 (*df*) = 8.443(2); CFI (comparative fit index) = .995; Bollen (IFI) fit index = .995; BBNNFI (Bentler-Bonnet non-normed fit index) = .984; Lisrel GFI (goodness-of-fit index) = .995; RMR (root mean square residual) = .013; RMSEA (root mean square error of approximation) = .063.

Good overall model fit was obtained on the confusion dimension for both NI and I South Africans. The results of the separate CFAs were also satisfactory, with CFI exceeding .95 and further fit statistics (BBNFI [Bentler-Bonnet normed fit index], NFI [normed fit index], RMSEA) indicating results beyond suggested threshold values (Bagozzi and Yi 1988).

This procedure was followed by invariance testing between corresponding paths of the two models, as suggested by Steenkamp and Baumgartner (1998). Configural invariance was established with a baseline model (unrestricted parameters) demonstrating a χ^2 (*df*) of 6.178(4), BBNFI = .995, CFI = .998, Lisrel GFI = .996, RMR = .011, RMSEA = .026. Fixing the four corresponding parameters to invariance (by introducing equality constraints) and using Lagrangian multiplier (LM) tests in EQS (Bentler and Wu 2002) to test whether the factorial structure was invariant across groups led to a significant drop in χ^2 (χ^2 = 38.407[4]). Thus, full metric invariance could not be established, although further testing indicated that partial invariance could be established.



Analysis Name RUNALL selected: Person Estimation by Weighted Maximum Likelihood method

Analysis Specifications

```
** ANALYSIS  TITLE       RUNALL

** SPECIAL  COMMENTS
         1.              Derived from the Default Project Settings

** TEST  STRUCTURE
Analysis Type            Polytomous/Extended Response Category test format
No. of Items             32
No. of Categories        Equal across Items : set at 5 categories per item
Score Range [All Items]  127
Some Items Anchored      No
Subtests created         No

** CALIBRATING  SAMPLE
No. of Persons:
  * entered Project      800
  * invalid records        0
  * extreme scores         0
  * valid scores         800 [available for analysis]
```

Analysis Name: runall

□ Project Default as the Basis

Select TEST type: ● Polytomous ○ Multiple-choice ○ Combined

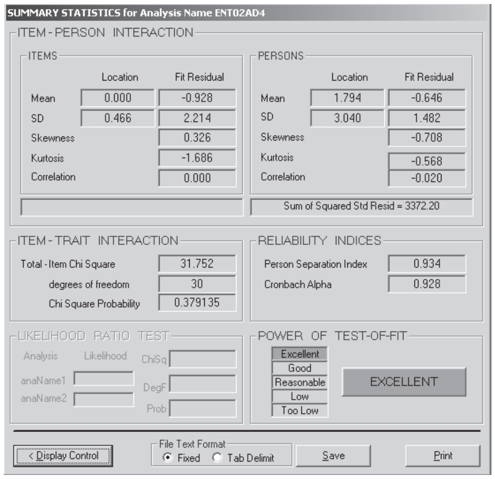
Analysis Options: ○ Run analysis ○ Edit analysis specifications ○ Create NEW analysis ● Display analysis outcomes ○ Targetted analysis

Modifications for New Analysis: ○ Analysis Base □ Delete sample: □ Delete items □ Likelihood-ratio Test ○ Subtest analysis ○ Absolute Anchoring ○ Relative Anchoring ○ Rescore items ○ Item Split [DIF analysis] ○ Edit Illegal Response Records

Test-of-Fit Control: ☑ Conduct Test-of-Fit; No. Class Intervals: 10; □ Include ANOVA Item Fit Statistics; Residual criterion: 2.0

Save Analysis Displays to ...: ○ **Single Batch File only** ● Individual Files to be specified

FIGURE E1
RUMM Project Definition



SUMMARY STATISTICS for Analysis Name ENT02AD4

ITEM - PERSON INTERACTION

ITEMS	Location	Fit Residual	PERSONS	Location	Fit Residual
Mean	0.000	-0.928	Mean	1.794	-0.646
SD	0.466	2.214	SD	3.040	1.482
Skewness		0.326	Skewness		-0.708
Kurtosis		-1.686	Kurtosis		-0.568
Correlation		0.000	Correlation		-0.020

Sum of Squared Std Resid = 3372.20

ITEM - TRAIT INTERACTION

Total - Item Chi Square	31.752
degrees of freedom	30
Chi Square Probability	0.379135

RELIABILITY INDICES

Person Separation Index	0.934
Cronbach Alpha	0.928

LIKELIHOOD RATIO TEST

POWER OF TEST-OF-FIT: EXCELLENT

File Text Format: ● Fixed ○ Tab Delimit

FIGURE E2
Summary Statistics

APPENDIX E

Illustrative Example of RUMM 2020 Software

This appendix aims to communicate some of the practicalities of equivalence testing using the Rasch methodology. It is hoped that this might also stimulate the adoption of the Rasch methodology within the advertising research community; consequently, the appendix provides screenshots of the software and annotations regarding some of the operational steps involved. This approach is designed to help quantitative researchers in capturing a fuller picture of the mechanics involved, even when not necessarily formally initiated to the latent trait theory.

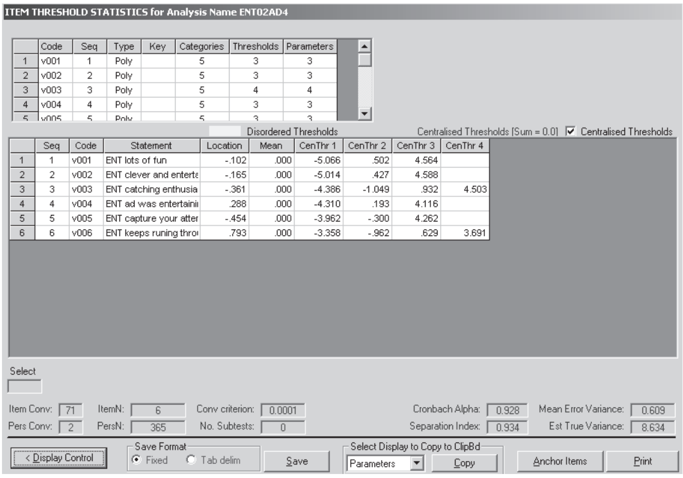
RUMM 2020 for Windows was used to perform the analyses in the paper. RUMM stands for "Rasch Unidimensional Measurement Models" and is a comprehensive item analysis package for the analyzing assessment and attitude questionnaire data. Analysis based on item response theory can also be performed using a recently developed SAS procedure (see Christensen and Bjorner 2003) or Linacre's (1991–2004) stand-alone "Winsteps" program. We decided to use RUMM, because it allows for user-friendly interaction procedures under the familiar Windows PC operating system and the software's methodological and scientific development is driven by acknowledged academic researchers (Andrich, Sheridan, and Luo 2003b).

Within RUMM, a project was defined based on the original data set (see Figure E.1). Within the project, analyses could be performed either comprising of the full data set or specific subsets of items or respondents. For each analysis, item scores could be changed to accommodate reversed thresholds. Similarly, items could be split according to some attribute of the respondents in order to account for DIF (differential item functioning) due to that attribute.

After running a Rasch analysis over a set of items, RUMM provides various output options, such as options for a graphical display of category probability curves, item characteristic curves (ICCs), or threshold probability curves. The software concludes the analysis with comprehensive summary statistics (see Figure E.2).

The summary statistics displays analysis results (pertaining to entertainment ads 1 and 3). The top left quadrant of the summary statistics screen shows the distribution of items. The mean of all item locations is set to zero by default. In this way, the origin of the scale is defined. The standard deviation is relatively small, indicating that the range of item locations is limited. The top right quadrant refers to respondents. Their mean location lies at 1.794, that is, the overlap of persons and items is not optimal. The item-trait interaction box on the bottom left of the summary statistics screen informs about interactions between individual respondents and items. This is essentially a general test of fit. To the right, reliability indices are reported. The person separation index (.934) is high, which is good. This index is a Rasch version of the reliability formula that takes Rasch measures and the standard errors into consideration. Low person separation results would be problematic, pointing at instruments that are of limited use or unable to differentiate between respondents. Consequently, this would also limit the power of the test-of-fit (see section at the bottom of the summary statistics screen).

RUMM further offers a screen displaying the individual item fit for the analysis. This reveals the location of the items, their standard errors, and fit indices (χ^2, F statistics, and respective probabilities). For polytomous items, threshold parameters are displayed in the item thresholds statistics window (see Figure E.3). Here, the centralized thresholds option is ticked, which means that the overall location of the item has not been taken into consideration but is reported separately. Hence, the threshold parameters add up to zero.



ITEM THRESHOLD STATISTICS for Analysis Name ENT02AD4

	Code	Seq	Type	Key	Categories	Thresholds	Parameters
1	v001	1	Poly		5	3	3
2	v002	2	Poly		5	3	3
3	v003	3	Poly		5	4	4
4	v004	4	Poly		5	3	3
5	v005	5	Poly		5	3	3

Disordered Thresholds   Centralised Thresholds [Sum = 0.0] ☑ Centralised Thresholds

	Seq	Code	Statement	Location	Mean	CenThr 1	CenThr 2	CenThr 3	CenThr 4
1	1	v001	ENT lots of fun	-.102	.000	-5.066	.502	4.564	
2	2	v002	ENT clever and enterta	-.165	.000	-5.014	.427	4.588	
3	3	v003	ENT catching enthusia	-.361	.000	-4.386	-1.049	.932	4.503
4	4	v004	ENT ad was entertainii	.288	.000	-4.310	.193	4.116	
5	5	v005	ENT capture your atter	-.454	.000	-3.962	-.300	4.262	
6	6	v006	ENT keeps runing throu	.793	.000	-3.358	-.962	.629	3.691

Select

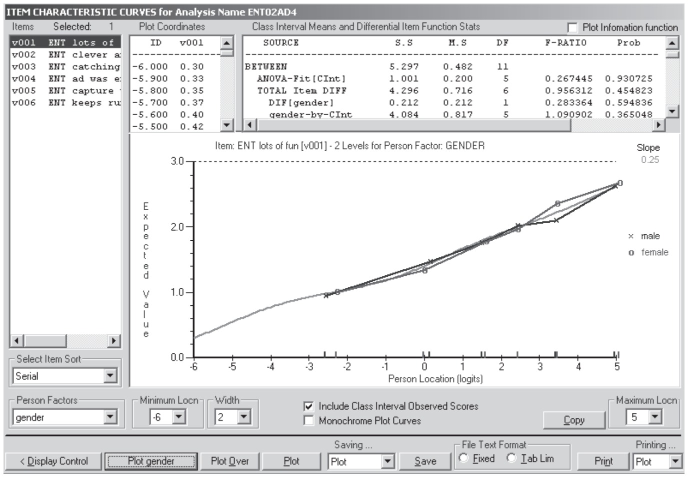
Item Conv: 71   ItemN: 6   Conv criterion: 0.0001   Cronbach Alpha: 0.928   Mean Error Variance: 0.609

Pers Conv: 2   PersN: 365   No. Subtests: 0   Separation Index: 0.934   Est True Variance: 8.634

Save Format: ● Fixed ○ Tab delim

Select Display to Copy to ClipBd: Parameters

FIGURE E3
Item Threshold Statistics



ITEM CHARACTERISTIC CURVES for Analysis Name ENT02AD4

Items   Selected: 1

```
v001  ENT lots of
v002  ENT clever a
v003  ENT catching
v004  ENT ad was e
v005  ENT capture
v006  ENT keeps ru
```

Plot Coordinates

```
  ID    v001
--------------
-6.000  0.30
-5.900  0.33
-5.800  0.35
-5.700  0.37
-5.600  0.40
-5.500  0.42
```

Class Interval Means and Differential Item Function Stats   □ Plot Infomation function

```
 SOURCE                S.S     M.S     DF     F-RATIO    Prob
-----------------------------------------------------------------
BETWEEN               5.297   0.482   11
  ANOVA-Fit[CInt]     1.001   0.200    5     0.267445  0.930725
  TOTAL Item DIFF     4.296   0.716    6     0.956312  0.454823
    DIF[gender]       0.212   0.212    1     0.283364  0.594836
     gender-by-CInt   4.084   0.817    5     1.090902  0.365048
```

Select Item Sort: Serial

Person Factors: gender   Minimum Locn: -6   Width: 2   Maximum Locn: 5

☑ Include Class Interval Observed Scores
□ Monochrome Plot Curves

Saving ...: Plot   File Text Format: ○ Fixed ○ Tab Lim   Printing ...: Plot

FIGURE E4
Item Characteristic Curves

For any respondent, reports can be retrieved that indicate this individuals' expected scores for any item, compared with the observed response. RUMM can also be requested to provide screens with person-item threshold distributions. These compare the distribution of respondents with the one from item thresholds. Although the distribution is not relevant for item calibration, it does matter for fit statistics and the precision of person and item estimates.

For a polytomous item, the item characteristic curve display (see Figure E.4) shows the expected value of the item score depending on the person location. For any person factor (e.g., culture or gender) included in the data sheet, separate actual values can be displayed and compared with each other. The result is a DIF analysis, the significance of which is assessed by an ANOVA (analysis of variance) test. In the example given, gender makes no difference at all, that is, the item means the same for males and females.