Journal of Business Research 66 (2013) 1307–1317

Contents lists available at SciVerse ScienceDirect

Journal of Business Research

Towards a new paradigm of measurement in marketing☆

Thomas Salzberger *, Monika Koller

WU Wien, Department of Marketing, Institute for Marketing Management, Augasse 2-6, 1090 Vienna, Austria

ARTICLE INFO

Article history:
Received 1 April 2011
Received in revised form 1 September 2011
Accepted 1 November 2011
Available online 10 March 2012

Keywords:
Measurement
Rasch model
Classical test theory
Item response theory
Formative indicator
Response instruction

ABSTRACT

In measuring latent variables, marketing research currently defines measurement as the assignment of numerals to objects. On this basis, marketing researchers utilize a multitude of avenues to measurement. However, the scientific concept of measurement requires the discovery of a specific structure in the data allowing for the inference of a quantitative latent variable. A review of the quite diverse approaches used to measure marketing constructs reveals serious limitations in terms of their suitability as measurement models. Adhering to a revised definition of measurement, this paper finds that the Rasch model is the most adequate available. An empirical example illustrates its application to a marketing scale. Furthermore, this study investigates how instructions about response speed and the direction of an agree–disagree response scale impact the fit of the data to the Rasch model and to confirmatory factor analysis. The findings are diametrically opposed, with Rasch suggesting more plausible conclusions. Rasch favors well-considered responses and the agree–disagree scale, while factor analysis supports spontaneous responses and the disagree–agree format. The adoption of the Rasch model as the foundation of measurement in marketing promises to promote more advanced and substantial construct theories, is likely to deliver better-substantiated measures, and will enhance the crucial link between content and construct validity.

1. Introduction

Testing theories about structural relationships between latent variables is the key objective of scientific market research (Howell, Breivik, & Wilcox, 2007). The analysis of such relationships relies on solid measurement models of the latent variables involved. Therefore, the issue of proper measurement is of utmost importance (Day & Montgomery, 1999) and represents a crucial field of enquiry (Lee & Hooley, 2005). Today, researchers utilize a multitude of approaches to measurement. Scientific marketing research is eclectic, regarding methodology (Creswell, 1994) in general and multivariate analysis techniques (Dahlstrom, Nygaard, & Crosno, 2008) in particular. A comprehensive range of methods is indeed advantageous in statistical analysis. However, the multiplicity of measurement approaches may actually be an obstacle to the advancement of quantitative research in marketing. It makes measurement appear to be a statistical task, involving fitting a model to a data set.

However, measuring latent variables is a scientific undertaking *sui generis*, which goes beyond a mere statistical problem (Michell, 1986, 1990). Specifically, a measurement model has to comply with the requirements of quantification. Therefore, model selection should not rest upon convenience, fit to the data (Andrich, 2004) or its extensive use in the discipline in the past (Stewart, 1981). This paper presents a scientific rationale for the requirements of measurement and then determines which measurement model best meets these requirements.

Firstly, a brief review of the approaches most widely used in marketing scrutinizes their suitability in tackling the challenge of measurement. Then, an empirical example, exploring the effects of response instructions and response scale direction, illustrates the applicability of an alternative measurement model in comparison to the traditional factor analytic approach.

2. Current approaches to measurement in marketing

2.1. Classical test theory

Measuring latent variables in marketing predominantly follows classical test theory (CTT) (Churchill, 1979; Lord & Novick, 1968). The fundamental idea of CTT is the separation of the true score from the error score, which add up to the observed score. The most popular CTT model is the congeneric model (Jöreskog, 1971), which applies this logic to the item level and explicitly accounts for the latent variable as the factor score.

However, the alleged explanation of an observed phenomenon using two unobservable components provides little insight and defies empirical rejection. Particularly, CTT fails to accomplish the goal of measurement, which is precisely what a measurement theory must

☆ The authors thank the JBR reviewers for reading and comments of earlier versions of this article.

* Corresponding author.
E-mail addresses: Thomas.Salzberger@wu.ac.at (T. Salzberger), Monika.Koller@wu.ac.at (M. Koller).

0148-2963/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbusres.2012.02.030

do. Rather, CTT treats scores as measures (Howell et al., 2007) and refers to aggregate statistics such as correlations and covariances, which presuppose measurement rather than justify quantification.

The linear relationship between the manifest score and the latent variable implies another shortcoming of CTT. While the span of the latent variable is theoretically infinite, the range of the manifest item score is limited to the number of response options provided. This potentially leads to floor and ceiling effects, when respondents hit the boundaries of the scale. In practice, researchers select items so that the mean respondent score is near the center of the scale and the distribution of the scores is close to normal (Likert, 1932) to avoid these effects.

As a result, all items typically characterize only a limited range of values of the latent variable, in terms of assessing how much of the property the items represent. Singh (2004) refers to this issue as the problem of narrow bandwidth. However, whether floor or ceiling effects occur also depends on the distribution of respondents. Hence, factor loadings and item reliability are contingent on the sample composition, even if the items are suitable indicators of the latent variable.

The sample dependence also entails that overall reliability confounds the properties of the instrument (error variance) and of the sample (true score variance). Hence, a low reliability need not imply a bad scale, especially when the sample is very homogeneous. In contrast, a high reliability may be a consequence of a heterogeneous sample, with many respondents hitting the floor or ceiling of the item scale. A general threshold for the reliability of a scale, such as the value of 0.7 recommended by Nunnally (1978), is therefore hard to defend.

Finally, a measure of an individual respondent or the mean of a group of respondents can only be interpreted in relation to measures of other respondents or the overall mean. In CTT, explicating a person's measure with reference to the items is impossible.

2.2. Formative models

CTT occasionally faces criticism in terms of the assumed flow of causality from the latent variable to the items called reflective indicators (Edwards & Bagozzi, 2000). Several authors suggest formative indicator models, or index models, where causality flows from the items to the latent variable, as an alternative to traditional scale development (e.g., Diamantopoulos & Winklhofer, 2001; Jarvis, MacKenzie, & Podsakoff, 2003; MacKenzie, 2003; Rossiter, 2002). In a formative model, the items define the construct. Since formative indicators represent different components of the construct, the index variable is typically multidimensional.

Therefore, formative indicators need not correlate, nor are they interchangeable. Validation procedures employing factor analysis and reliability estimation based on internal consistency are therefore inappropriate. Besides content validity, external validity plays an important role (see Diamantopoulos & Winklhofer, 2001).

Procedures common in CTT are inapplicable because formative models do not deal with a measurement problem per se. Index variables are composite variables, often misinterpreted as latent variables (Stenner, Burdick, & Stone, 2008; Stenner, Stone, & Burdick, 2009; Ping, 2004). An index summarizes multiple measures (Borsboom, 2005) and, as such, presumes the measurement of all its components. Unfortunately, papers on formative models typically obscure the crucial difference between measurement and summarizing measurements. The only case where formative models do deal with measurement is the multiple indicators multiple causes model (MIMIC, Edwards & Bagozzi, 2000; Jöreskog & Goldberger, 1975), which features both reflective and formative indicators. However, the relationship between the formative indicators and the latent variable actually constitutes a structural model, whereas the reflective indicators form the measurement model (Howell et al., 2007; Wilcox, Howell, & Breivik, 2008).

According to Jarvis et al. (2003, p.203) the condition that "changes in the construct are not expected to cause changes in the indicators" argues for a formative model. The notion of indicators that actually do not indicate a change in the construct demonstrates that formative indicators are not concerned with the problem of measurement. Thus, formative and reflective models do involve the potential for misspecification. However, the misspecification does not take place within the sphere of measurement but between a measurement model and a structural model. Consequently, the formative model is not a measurement model (Ping, 2004), notwithstanding its potential usefulness as a summary of measures, for example for the purpose of prediction.

2.3. The C–OAR–SE approach

C-OAR-SE stands for construct definition–object representation–attribute classification–rater entity identification–selection of item type and answer scale–enumeration and scoring rule (Rossiter, 2010, p.2). The approach (Rossiter, 2002, 2010), which sets out to replace psychometrics, reveals many of the shortcomings of CTT. Psychometrics, according to Rossiter (2010), merely model the relationship between measures and scores, relying excessively on statistics to assess validity, while ignoring the construct and disregarding content validity. In C-OAR-SE, the construct definition includes the object to which the attribute refers (e.g., a particular company), and the rater entity (e.g., consumers), as well as the attribute itself (e.g., service quality).

Diamantopoulos (2005) criticizes this definition, but Rossiter's argument deserves close attention. The interpretation of measures across different objects and/or different raters in traditional measurement typically treats the differences as if they do not matter. Rossiter's view is diametrically opposed to this belief, as different objects (or raters) imply different constructs. However, an empirical science could, and should, make this an empirical question and seek appropriate evidence. The hallmark of C-OAR-SE though is its reliance on content validity, that is expert validation, as the only essential type of validity. Reminding us of the importance of content validity is a merit indeed. However, C-OAR-SE abandons statistical analysis and thereby decouples the task of measurement from the empirical evidence (Finn & Kayande, 2005). Hence, C-OAR-SE captures the essence of measurement incompletely and lacks the requirements of a measurement theory. Nevertheless, C-OAR-SE, interpreted as an understandable counter-reaction to the overreliance on sometimes doubtful statistical measurement models, vividly reminds us of the importance of content validity.

2.4. Item response theory

In contrast to CTT, item response theory (IRT; Hambleton, Swaminathan, & Rogers, 1991) focuses on individual responses to particular items rather than aggregate statistics. All IRT models specify a respondent location parameter, which is the measure of ultimate interest, and a set of item parameters, typically a location and a discrimination parameter. The item location parameter specifies the amount of the property the item accounts for. A logistic, s-shaped function models the relationship between the respondent location and the response probability, given the item properties.

The item characteristic curve (ICC) depicts this function graphically. The item discrimination parameter determines the slope of the ICC. The logistic function accommodates the fact that the manifest score is restricted to a small number of responses, bounded between a minimum and a maximum. With polytomous items, rather than modeling the item score, each individual response category is considered explicitly, thus accounting for the discrete character of the responses.

IRT models differ in terms of the parametrization of the items. The Birnbaum model (1968) features a discrimination parameter for each item. In contrast, the Rasch model requires item discrimination to be equal across items. This difference has important theoretical and philosophical consequences. Specifically, the Rasch model has unique properties, with specific objectivity (Rasch, 1977) as its defining characteristic. Thus, Rasch models are a separate class to be distinguished from the general IRT models (see Andrich, 2004). While general IRT seeks an optimal description of the data, in Rasch measurement the model takes precedence over the data. Where the data do not fit the Rasch model, IRT proponents resort to a more general model, whereas advocates of the Rasch model take this to indicate serious problems, rejecting the hypothesis of measurement.

Singh (2004) presents an introduction to the use of IRT as a measurement model in marketing, focusing on the Birnbaum model (1968) for dichotomous responses. Applications of Rasch measurement in marketing are still scarce but their number appears to be increasing (e.g., Bechtel, 1985; Bechtel & Wiley, 1983; Ganglmair & Lawson, 2003a, 2003b; Ewing, Salzberger, & Sinkovics, 2005; Salzberger & Sinkovics, 2006).

2.5. Conclusion

The approaches to measurement in marketing differ widely. CTT focuses on correlation statistics, reliability and factorial validity, with causality flowing from the latent variable to the items. Formative models reverse the causality and use content validity as their prime justification but also refer to external validity. The C-OAR-SE approach neglects empirical evidence based on respondent data, focusing on content validity and expert judgment. IRT models the individual item response as a non-linear function of the latent variable and item properties. However, general IRT and the Rasch model hold very different views as to the precedence of the data versus the model. The question arises: how is it possible that a range of vastly different approaches can all purport to yield interval scale measures of latent variables? The answer lies in the definition of measurement in the social sciences.

3. Definition of measurement

3.1. Current definition of measurement in marketing

The current definition goes back to Stevens (1946, 1951), according to whom “[m]easurement is the assignment of numerals to objects or events according to rules” (1951, p.1). Interestingly, the definition typically remains implicit. Even methodological contributions being critical of formative measurement models (e.g., Wilcox et al., 2008) remarkably do without an explicit definition of measurement. This fact implies that Stevens' definition is effectively unchallenged.

Stevens never specified the rules exactly, merely pointing out that they have to be consistent (Stevens, 1959). However, the assignment of numerals to objects or events simply describes the process of coding. Coding is essential as it summarizes observations but measurement aims to infer a quantitative latent variable from these coded observations. All of the approaches discussed are compatible with Stevens's non-committal definition. More than 30 years ago, Jacoby (1978, p.91) pointed out that “most of our measures are only measures because someone says that they are, not because they have been shown to satisfy standard criteria”. Stevens's definition provides no guidance as to the sort of empirical evidence required. However, the type of support informs the proper choice of a measurement model.

3.2. The role of the measurement model

The definition of a construct is a vivid description of what the researcher has in mind. Although a definition cannot be true or false as such, the construct that is defined may or may not have a counterpart in the real world. A suggested construct implies a hypothesized latent variable, an ontological claim (Borsboom, 2005), which is subject to empirical corroboration. Researchers typically imply that the suggested and measured latent variable exists independently of the measurement. Thurstone (1928, p.532), for example, speaks of “real attitude”. Borsboom (2005, p.58) points out that the realist interpretation is the only view that “can be consistently attached to the formal content of latent variable theory”.

With structural relationships between latent variables, the empirical analysis has to corroborate the theory. The same principle applies to a hypothesized latent variable. The measurement model has to provide empirical criteria which address the ontological claim and therefore allow for tentatively accepting the construct as real. Rossiter (2010) is correct when objecting to measures that derive their justification solely from the application of a psychometric model. Indeed, content validity is indispensable, as is construct validity. The assessment of construct validity has to account for the empirical consequences of content validity and the requirements of quantity.

Most so-called measurement theories fail to incorporate the requirements of quantity. C-OAR-SE (Rossiter, 2010, p.13) rejects construct validity altogether. Formative models are incompatible with the “formal content of latent variable theory” (Borsboom, 2005, p.58). CTT is concerned with a decomposition of variance, which can be useful but presumes rather than constitutes measurement. In contrast, IRT seems to be the most promising candidate as it provides a model of how measurement can be accomplished. However, only a revised definition of measurement would allow a decision in favor of a particular IRT model.

3.3. A revised definition of measurement

Stevens's definition of measurement is a reaction to the apparent inability of psychology to qualify as a quantitative science based on the definition of measurement in the natural sciences (see Michell, 1986, 1999). By adopting Stevens's definition, the social sciences, first, disconnected quantification and measurement and, second, contributed to the separation of social from natural sciences. The concept of measurement can only regain its rigor in the social sciences by returning to the uniform definition of measurement of the natural sciences. In the latter, measurement “is the process of discovering ratios” (Michell, 1999, p.14) that are independent of the measurement unit. For example, when the measurement of length reveals that two objects are in a ratio of 3:1, this holds true regardless of the unit. The measurement of an attribute of an object “involves comparisons with a unit” (Michell, 1999, p.186) and “[t]he unit of measurement on any scale for any quantitative attribute … is that magnitude of the attribute whose measure is 1” (Michell, 1999, p.13).

Michell (1997, 1999, 2003) distinguishes between two tasks of measurement. The scientific task, concerned with evidence that the latent variable actually is quantitative, tackles the ontological claim. The instrumental task deals with the actual implementation of measurement. Measurement based on assignment neglects the scientific task, which “sounds a warning to us about the nature of measurement in the social sciences” (Balnaves & Caputi, 2001, p.51). The scientific task of measurement consists of showing that the measures sustain additivity. Properties such as length allow for a direct empirical demonstration of additivity using concatenation operations (Michell, 1990). In the social sciences, evidence of additivity is indirect, similar to derived measurement in the natural sciences. Density, for example, is a function of mass and volume. The discovery of constants that emerge when an increase in mass by a particular ratio is counterbalanced by an increase in volume by the same ratio, provides evidence that density is quantitative (Michell, 1999).The same principle is applicable in the social sciences (Luce & Tukey, 1964).

3.4. Axioms of measurement

The axiomatic framework of measurement, outlined by Hölder (1901) and translated into English by Michell and Ernst (1996, 1997), permits the derivation of the required features of a measurement model. Specifically, the axioms imply so-called cancelation conditions (Karabatsos, 2001), which refer to order relations between the elements in the data matrix. The probabilities of a response, implied by the measurement model, should comply with these conditions. Single cancelation requires, for example, that if a respondent has a higher probability of agreeing with item v than with item w (one premise), then this has to be true for all respondents (one consequence). Double cancelation entails two premises and one consequence (see Karabatsos, 2001, for details). The cancelation conditions force the ICCs to be parallel. Since general IRT models allow for intersecting ICCs due to unequal item discrimination, they are incompatible with the axioms. In contrast, Rasch models do comply with the cancelation conditions.

3.5. Specific objectivity in the Rasch model

Specific objectivity (Rasch, 1977) is the defining feature of the Rasch model. This principle implies the requirement of invariance, which provides a more intuitive avenue to the Rasch model. In practical terms invariance means that person measures have to be independent of the items researchers actually use (provided all items are suitable for measuring the same latent variable) and vice versa (the item measures have to be independent of the people). Invariance is a property of the model as much as it is of the data. The Rasch model builds upon invariance but empirical data may falsify this requirement. The extent to which the invariance is given determines the boundaries of the frame of reference within which the measures are meaningfully comparable.

4. The Rasch measurement approach

The Rasch model accommodates dichotomous responses (dichotomous Rasch model, Rasch, 1960) as well as responses to polytomous items (Andrich, 1978a, 1978b, 1988; Masters, 1982). In the dichotomous case, the difference between the person location parameter β_v and the item location parameter δ_i governs the probability of a positive (affirmative) response by respondent v to item i (Eq. (1)). Item discrimination is implicitly set at 1 for all items.

$$P(a_{vi}=1)=\frac{e^{(\beta_v-\delta_i)}}{1+e^{(\beta_v-\delta_i)}} \tag{1}$$

For the estimation of the person and item location parameters, the conditional maximum likelihood estimation (CML; Molenaar, 1995) and the pairwise estimation approach (Andrich & Luo, 2003; Zwindermann, 1995) lend themselves particularly well as they maintain the invariance property. In contrast to CTT and general IRT, the Rasch model does not require any assumptions about the distribution of the person parameters.

The polytomous Rasch model specifies a set of threshold parameters (τ_{ij}), which characterize the response categories of a polytomous item. τ_{ij} indicates the transition point between two adjacent response categories, j and j + 1, within item i, thus representing the point where responses to two adjacent categories are equally likely. The average of the m − 1 threshold parameters for an item with m response categories represents the overall location (δ_i) of that item.

The empirical threshold estimates are not restricted at all and can thus be in any order (Andrich, 1995a, 1995b). Disordered threshold estimates indicate that the response categories fail to function as intended. Eq. (2) shows the probability of choosing a category scored x (ranging from 0 to m), given the person location β_v, the item overall location δ_i and the threshold locations τ_{ij} (Andrich, 1988). The denominator γ is the sum of the numerators across all response categories. The numerator for the lowest category (0) is 1. In the polytomous case, the ICC describes the expected item score as a function of the person location, while the category characteristic curves depict the probability of each category.

$$P\left(a_{vi}=x|\beta_v,\tau_{ij},j=1\ldots m,0<x\le m\right)=\frac{e^{\left(\sum_{j=1}^{x}-\tau_{ij}\right)+x\cdot(\beta_v-\delta_i)}}{\gamma} \tag{2}$$

with,

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

Besides parameter estimation, testing the fit of the data to the Rasch model is crucial and is a comprehensive task, with no single routine capturing all possible occurrences of lack of fit. A visual comparison of the empirical ICC, representing the actual mean scores of groups of homogeneous respondents (dividing the continuum into several classes), with that expected under the model provides a first insight into the functioning of the item. A chi-square test of fit provides statistical evidence. Other statistics, such as the fit residual statistic in RUMM 2030 (Andrich, Sheridan, & Luo, 2010), highlight under- or overdiscrimination of the actual item. In the case of polytomous items, an investigation of the order of the threshold parameters should accompany the interpretation of fit statistics. Ignoring reversed thresholds invalidates the scoring of successive categories.

The invariance property offers another avenue for testing fit. A likelihood-ratio test assesses whether item calibrations from different subsets of respondents are equal up to random variation. A test of item differential functioning (DIF) investigates the same principle at the item level. DIF means that an item works differently for different subgroups, defined by, for example, gender, culture, or other personal factors. Ideally, a measurement instrument is free of DIF. However, the retention of a DIF-affected item is possible, provided invariant items define the metric of the latent variable unequivocally. In this case, the estimation of location parameters for DIF items occurs separately for different groups.

Local independence, that is the lack of local dependence (LD), is another important criterion of fit. LD means that two items remain related after accounting for the latent variable. In most cases, LD is due to multidimensionality or response dependency. Response dependency occurs when agreeing with one item logically entails agreeing with another. The analysis of residual correlations allows for an assessment of LD. Values outside the interval of −0.2 and +0.2 require attention. Combining items into a super-item, or omitting one, may solve LD in this case. If multidimensionality prevails, the latent variables should be disentangled. A principal component analysis (PCA) of the item residuals helps to determine whether unidimensionality is questionable. The eigenvalues of the PCA should not deviate from a random pattern, as determined, for example, by parallel analysis (Allan & Hubbard, 1986; Horn, 1965; Watkins, 2000).

No unique path assesses fit as one kind of problem in the data may affect another. The power of the test of fit is another relevant issue. Besides sample size, the targeting of the instrument is the main determinant of power. When the person distribution grossly mismatches the distribution of item locations, the items do not discriminate between the respondents and vice versa. The person-separation-index (PSI; Andrich, 1982) gives the proportion of true variance in the person estimates over total variance, which includes error variance. The PSI corresponds to the coefficient alpha in CTT.

The validation of a measurement instrument should encompass more than just quantitative fit statistics. The fit of the data to the Rasch model is a requirement but does not necessarily imply that the hypothesis of the latent variable receives support. Since the Rasch model builds on the variation of the items, in terms of the amount of the property they represent, it is possible to establish a link between content validity and construct validity. Ultimately, the order of the empirical item locations should correspond to the expected order according to the conceptual definition of the construct (Wilson, 2005).

5. Empirical example

5.1. Function of the empirical example

The empirical example illustrates the applicability of the Rasch model to an existing scale in marketing, compared to CTT, but does not aim to provide evidence that the Rasch model is superior since the suitability of a measurement model is a theoretical problem. Empirical exemplifications are incapable of justifying the adequacy of a measurement model. This philosophy is in sharp contrast to the traditional view, according to which a multitude of models may serve the purpose of measurement. However, the revised definition, which the natural sciences have always used, forces researchers to grant precedence to theory over data.

In addition, the study investigates the psychometric effects of differently directed response scales and response instructions in an experimental setting.

5.2. New Environmental Paradigm (NEP) scale

Given the pressing problems related to environmental issues, the effect of overuse of resources is a more topical issue than ever in consumer behavior research (Griskevicius, Tybur, & Van den Bergh, 2010; Kilbourne, Beckmann, & Thelen, 2002). In 1978, Dunlap and Van Liere (1978) proposed the New Environmental Paradigm Scale (NEP), comprising 12 items. The theory behind the scale is built on a literature review as well as input from environmental scientists and ecologists. Later, Dunlap, Van Liere, Mertig, and Jones (2000) suggested a revised 15-item NEP scale (now signifying New Ecological Paradigm) without greatly altering the construct definition. While omitting four items from the original, the revised instrument features seven new ones. Research on pro-environmental orientation has frequently utilized the original scale (Dunlap et al., 2000). In terms of dimensionality, empirical evidence based on factor analysis is quite mixed. While some studies report multiple dimensions, Dunlap et al. (2000) recommend combining all items into a single measure.

The present example uses the original NEP scale (see Table 1) as, from a Rasch perspective, the new items seem to be more problematic than the original ones. However, future empirical research has to examine how the new items blend in.

Table 1
12 NEP items administered in the questionnaire (Dunlap & Van Liere, 1978).

Nep 1	–We are approaching the limit of the number of people the earth can support. [misfitting]
Nep 2	–The balance of nature is very delicate and easily upset.
Nep 3	–Humans have the right to modify the natural environment to suit their needs.
Nep 4	–Mankind was created to rule over the rest of nature. [misfitting]
Nep 5	–When humans interfere with nature it often produces disastrous consequences.
Nep 6	–Plants and animals exist primarily to be used by humans.
Nep 7	–To maintain a healthy economy we will have to develop a “steady-state” economy where industrial growth is controlled. [misfitting]
Nep 8	–Humans must live in harmony with nature in order to survive.
Nep 9	–The earth is like a spaceship with only limited room and resources.
Nep 10	–Humans need not adapt to the natural environment because they can remake it to suit their needs.
Nep 11	–There are limits to growth beyond which our industrialized society cannot expand.
Nep 12	–Mankind is severely abusing the environment.

Another reason for using a scale measuring environmental attitudes is its potential susceptibility to socially desirable responses, which is important given the interest in the effects of different response instructions.

5.3. Experimental setting

The scale presented to the respondents uses a three times two experimental design giving six conditions. The first factor provides three types of response instruction; the second two types of response scale. A random allocation of respondents ensures internal validity.

5.3.1. Response instructions

While many published marketing scales do not specify a desired speed of response, some ask for spontaneous responses or, less often, explicitly invoke a well-thought-out response. The Imagery Vividness scale (Childers & Heckler, 1985) instructs respondents to “consider carefully the picture that comes before your mind's eye” (Bruner & Hensel, 1996, p.321). Examples of scales asking for spontaneous responses include the Novelty Experience Seeking scale (Bruner & Hensel, 1996, p. 454; Venkatraman, 1991; Venkatraman & Price, 1990), the Motivation to Work scale (Bruner & Hensel, 1996, p. 967), and the Perception Survey for Health-Care Workers scale (WHO, 2009). Practical guidelines to questionnaire design often mention the lack of spontaneous responses as a drawback. Nimkar (2010) advises questionnaire developers to “[a]sk [the] respondent to give [a] spontaneous response to know his top of mind awareness”.

Empirical studies on the effects of response instructions are scarce in marketing research but experimental research in psychology contributes a few such studies (e.g., Lievens, Sackett, & Buyse, 2009; McDaniel, Hartman, Whetzel, & Lee Grubb III, 2007; Wenke & Frensch, 2005). Instructions about speed usually aim to elicit impulsive versus considered answers. Impulsive answers are thought to mirror more intuitive, emotional and subconscious attitudes, while considered ones more cognitive mental processing. Time constraints also impact on the decision-making strategies of respondents (Glöckner & Betsch, 2008).

In the present study, the instructions ask the first group of respondents to carefully consider their responses and think about the pros and cons before choosing a response category and the second group to reply spontaneously without weighing their reaction. Respondents in the third, control group receive no instructions as to response speed.

5.3.2. Response scale direction

The interpretation of the item, the retrieval of relevant beliefs and feelings, and resulting judgment determine the qualitative response (Tourangeau & Rasinski, 1988). The optimal design of the response scale facilitates the transformation of the qualitative into a quantitative response. The literature contains a plethora of studies investigating design options for response scales, for example the appropriate number of response categories (e.g., Weng, 2004), using an odd or even number of categories, the spacing between categories, verbally or numerically labeling response options, and direction of response scale, that is, agreement to the left and disagreement to the right (the A≫D format) or the reverse (D≫A).

Evidence in the literature as to which direction performs better is inconclusive. Dickson and Albaum (1975) find no mean differences with semantic differential scales, while Holmes (1974) and Mathews (1929) report a bias towards the left. Sheluga, Jacoby, and Major (1978) find stronger agreement in the A≫D format, but no differences in terms of variances and reliability. Friedman, Friedman, and Gluck (1988) and Friedman, Herskovitz, and Pollack (1994) confirm this additive bias in college evaluation data. In contrast, Weng and Cheng (2000) find no relevant differences in terms of means, inter-item correlations and covariances, and the factor structure. Bradburn, Sudman, and

Wansink (2004, p. 161) conclude that there is “no good evidence that one form is universally better than another”, as multiple factors apparently interact.

In the present study, half of the respondents have to choose between seven response categories in the A ≫ D format. The remaining respondents use the reversed format (D ≫ A). In either case the statement is presented on the left hand side and the response scale on the right.

5.4. Hypotheses

Well-considered responses can be of higher quality in multiple ways. First, the respondent spends more time interpreting the item and retrieving relevant beliefs and feelings. Second, asking respondents to consider their responses encourages them to access their beliefs and feelings more comprehensively. Thus, the judgments should be more valid and the consistency between observed responses and the latent variable enhanced. Hypothesis A therefore proposes that the measurement model fits best when the respondents consider their answers carefully.

On the other hand, well-thought-out responses also increase the chance of social desirability playing a role, because respondents have more time to consider what a socially acceptable response might be. In fact, researchers typically request spontaneous responses as a means of reducing susceptibility to socially desirable responses. Accordingly, hypothesis B states that the mean of the group asked to give well-considered responses will indicate a more pro-environmental attitude than that of the spontaneous group.

In terms of the response scale direction, it is much harder to establish a clear theoretical advantage of one format over the other. The D ≫ A format is compatible with the usual arrangement of scales in graphical depictions starting at a low (or zero) amount to the left and progressing towards a higher amount to the right. Hypothesis C1 therefore assumes that the D ≫ A version of the response scale works better than the A ≫ D format. However, whether or not the response process actually follows this rationale is an issue. Tourangeau, Couper, and Conrad (2004) distinguish between five interpretive heuristics based on the concepts of Gestalt theory, one of which is of particular relevance to the issue at hand. The heuristic “near means close” (p. 370) suggests that respondents interpret stimuli which are presented near to each other as more closely related than those that are spatially separated. This heuristic may also apply to the spatial relationship between the item (the statement) and the associated response scale. Thus, spatial proximity between the statement and the agree-pole of a Likert-type response scale should be more intuitive than proximity to the disagree-pole. This argument leads to hypothesis C2, which expects the A ≫ D version to better fit the measurement model.

5.5. Data set

The data set comprises a total of 1644 respondents. To collect data, the researchers administered a questionnaire online among students at a major European Business School. The sample is relatively homogeneous in terms of age (mean 25 years, 97% are under 41) and certainly does not represent the entire population. However, since the purpose of the study is not to generalize to the wider population, this restriction is not a severe limitation. Moreover, the Rasch model does not depend on representative samples to the same extent as CTT does. Provided that the data fit the model, the item parameters are independent of the sample composition due to the invariance property of the Rasch model. Tests for differential item functioning, for example based on gender, check the actual sample invariance in the data. However, this approach is limited inasmuch as sample independence cannot extend beyond the characteristics of the respondents actually present in the sample. Thus, respondents over the age of 40 are essentially outside the frame of reference of the current example.

5.6. Scale analysis using the Rasch model

RUMM 2030 (Andrich et al., 2010) provides parameter estimations and fit assessments using the extended Rasch model for polytomous items with unconstrained threshold estimates (Andrich, 1988) and the rating scale model (RSM; Andrich, 1978a), which specifies a uniform threshold structure across all items. The initial scale analysis focuses on the sample of respondents in the control group (no response speed instructions) using the A ≫ D version of the response scale. This group consists of 273 respondents, which is sufficiently large to assess the item properties appropriately (Ewing et al., 2005).

The analysis proceeds iteratively, that is only one deletion of an item occurs at a time. Eventually, nine items fit the RSM very well. This item set also represents the best fitting analysis within each experimental condition. However, clear differences in terms of overall fit occur across conditions. The person-separation index (PSI) ranges between 0.70 and 0.78 for all analyses. Therefore, the power of the test of fit is satisfactory and roughly equal across conditions with essentially the same sample size.

Two items, 7 and 1 (see Table 1), show a clear lack of fit in terms of the comparison of expected and actual responses eventually resulting in their omission from further analysis. Item 7 works reasonably well for respondents scoring low, with a steady increase in the empirical ICC determined by the dots representing actual mean responses (see Fig. 1). However, most respondents react uniformly to the item. In fact, the wording of the item (see Table 1) may be too sophisticated. The meaning of “controlled industrial growth” leaves some wiggle room. Item 1, “We are approaching the limit of the number of people the earth can support,” appears to lack clear ties to the key construct, whether humans should dominate nature or seek harmony.

Deleting items 7 and 1 leads to a satisfactory fit. However, the investigation of the item residuals reveals noteworthy correlations involving items 4 and 3 ($r = 0.29$), and 4 and 5 ($r = 0.24$), indicating response dependency. The removal of item 4 solves this redundancy. In terms of unidimensionality, the eigenvalue of the first principal component amounts to 1.886, while a parallel analysis (Monte Carlo PCA, Watkins, 2000) gives 1.29. The loadings on the first principal component indicate that items 3, 6 and 10 may depict another dimension, which relates to mankind having the right to dominate the environment.

A comparison of respondent measures based on items 3, 6 and 10 and measures based on the remaining items shows that only 6.96% of all respondents have significantly different estimates. Since the confidence interval for proportions, that is 6.96% ± 3.00%, includes 5%, which is the chance level of significant differences at $\alpha = 0.05$ given unidimensionality (see Smith, 2002, for details), unidimensionality is tenable. For most conditions, the threshold estimates are properly ordered. Any disordering is marginal and appears to be accidental as it occurs at the lower end of the scale where respondents are sparse and information is rather limited.

Finally, a DIF analysis based on gender shows that all items work equivalently for males and females. The targeting of the scale is somewhat limited when applied to students. Their person mean measure is 0.8, which is slightly above the item mean location of 0.0 (see Fig. 2, on which respondents are plotted upwards and items downwards). A limitation lies in the relatively small variation in the item locations, which reduces precision for respondents scoring very low or very high. This shortcoming is not surprising though, as the scale development adhered to classical standards, where variation in item locations is not an issue.

Consequently, a set of nine NEP items performs very well overall with an excellent fit of the data to the RSM (χ^2 (81) = 90.3, $p = 0.225$; see Table 2 for item characteristics and fit indices).

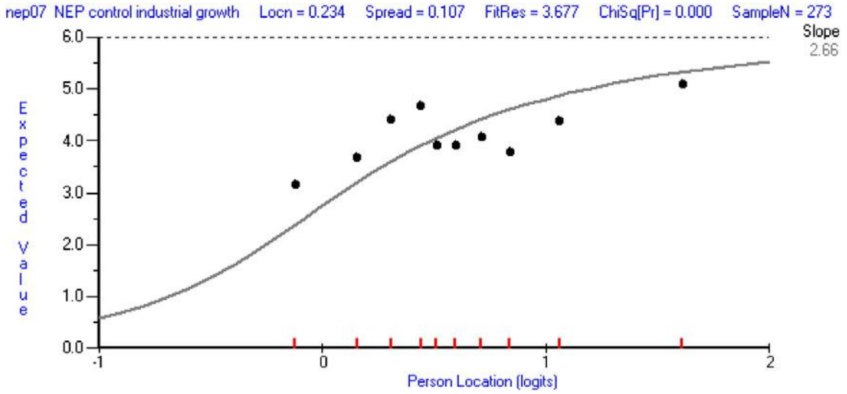

Fig. 1. The expected ICC under the model and actual responses in 10 classes of respondents.

5.7. Experimental conditions

The first analysis (analysis 1, Table 3) refers to the entire sample of 1644 respondents, and all six experimental conditions. The extraordinarily bad overall fit suggests heterogeneity across experimental conditions. Analyses 2 and 3 confine the data set to the A ≫ D and D ≫ A response scales, respectively. The overall fit remains unsatisfactory in either case. Analyses 4 to 6 refer to homogeneous subsets of the data in terms of response instructions. These analyses deliver a clear indication that the instruction to respond in a well-considered way performs best, with the control group ranking a close second. The instruction to give spontaneous responses results in the worst fitting data.

Since the experimental conditions might interact with each other, analyses 7 to 12 apply to each combination of response instruction and response scale direction. Within the well-considered instruction group and the control group, the A ≫ D scale is superior to D ≫ A. No such difference occurs in the spontaneous response group. This suggests a main effect for the scale direction, favoring hypothesis C2 and rejecting hypothesis C1 (see Fig. 3).

Likewise, use of the instruction to give well-considered responses performs best, while use of the instruction to respond spontaneously impacts negatively on the fit. However, the differences between the well-considered responses and those in the control group are small. This finding suggests another main effect supporting hypothesis A (well-considered responses provide a better fit than spontaneous ones).

Besides the main effects, an interaction effect occurs, as there is no difference between the A ≫ D and the D ≫ A scale when the instruction prompts respondents to respond spontaneously. Fig. 3 visualizes the main effects and interaction effect.

Since respondents do not necessarily comply with response instructions, the analysis takes the actual response speed into consideration as well (analyses 13 to 18; the 33rd and the 67th percentile of the response speed distribution define the three groups). It turns out that a fast response process leads to a rather bad fit. The same applies to a moderate speed when in conjunction with the D ≫ A response scale. A moderate response speed when using the A ≫ D scale provides the best fit, with slow responses providing a good fit too. Interestingly, a slow response process seems to allow the respondent to cope with the generally inferior D ≫ A scale. Since the PSI is slightly higher in the case of a high response speed, the differences in fit might partly be due to variations in the power of the test of fit. Moreover, individual respondents vary in terms of what constitutes a high response speed for them. In any case, the analyses based on actual speeds confirm the conclusions about response instructions.

Finally, no mean differences occur across any of the investigated conditions. Thus, the study provides no indication that social desirability is more of an influence on well-considered responses or that spontaneous response prevents respondents from considering the issue. Thus, hypothesis B does not receive confirmation.

In summary, the results support hypothesis A (well-considered responses do yield better fitting data), while rejecting hypothesis B (spontaneous responses do not suppress social desirability). In terms of the response format, data based on the A ≫ D scale fit better than the reverse D ≫ A, favoring hypothesis C2 and rejecting C1.

5.8. A classical test theory perspective

The analysis of the same data set based on confirmatory factor analysis (CFA, maximum likelihood estimation, carried out using Mplus, developed by Muthén & Muthén, 2008) aims at comparing the results from the Rasch analysis with CTT. The focus is on the relative differences in the fits of different experimental conditions rather than an absolute fit assessment. The appraisal of fit rests on the χ^2 value and the root mean square error of approximation (RSMEA;

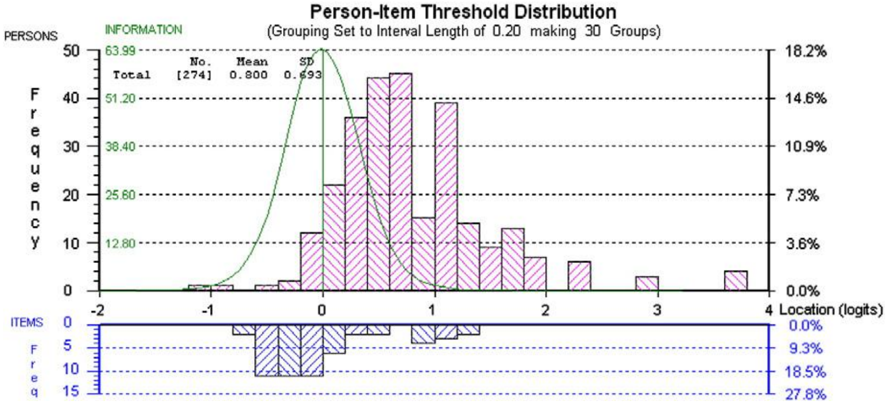

Fig. 2. Targeting plot for analysis 5 (control group, A ≫ D response scale).

Table 2
Item characteristics and fit statistics (analysis 5).

Item	Location	SE	FitResid	ChiSq	DF	Prob
Nep02	0.06	0.050	0.90	12.43	9	0.19
Nep03	0.27	0.05	1.42	7.36	9	0.60
Nep05	−0.17	0.06	3.36	10.33	9	0.32
Nep06	0.02	0.05	1.99	8.85	9	0.45
Nep08	−0.20	0.06	−1.76	9.86	9	0.36
Nep09	0.12	0.05	−0.47	7.41	9	0.59
Nep10	−0.20	0.06	0.76	18.43	9	0.03
Nep11	0.23	0.05	0.43	6.25	9	0.71
Nep12	−0.13	0.06	−0.97	9.37	9	0.40

Browne & Cudeck, 1993). However, other fit indicators, including the sample-size adjusted BIC, support the findings as well. The same applies to additional estimations treating the variables as categorical and using the weighted least squares estimator (WLSMV) in Mplus (Table 4).

As in the Rasch analysis, analysis 1 comprises the whole sample and fits very badly compared to all other models estimated. The more homogeneous conditions of analyses 2 and 3 perform better. Interestingly, in the CFA, the D≫A scale (analysis 3) seems to work better than A≫D (analysis 2). The same occurs in relation to the response instructions. According to the CFA analyses, the spontaneous response process (analysis 4) yields the best fitting CFA model, and the well-considered (analysis 5) the worst.

Analyses 7 to 12 (see Fig. 4) confirm the main effect that D≫A performs better on average. Another main effect shows that the spontaneous instruction works best, while the well-considered one functions worst. As in the Rasch findings, an interaction effect seems to occur although it is harder to interpret. Using the D≫A format, the instruction has only a minor impact, with the control group performing worst, which defies straightforward interpretation. The A≫D version works better than the D≫A version when used with the instruction to respond spontaneously and in the control group, but worse in the well-considered group.

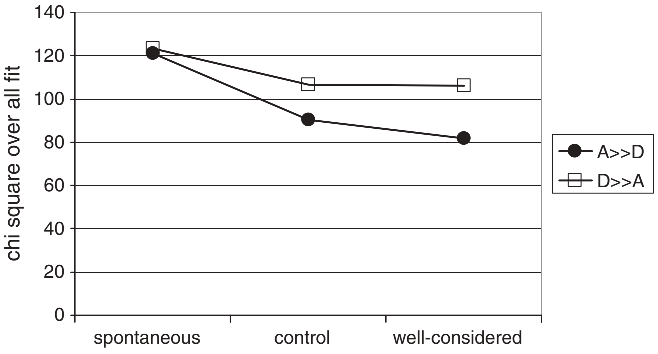

Fig. 3. Chi-square fit statistics for all six experimental conditions (Rasch model).

5.9. Comparing Rasch and CFA findings

The divergence of the results of the Rasch analysis and the conventional CFA raises questions as to which findings are more trustworthy and the reasons for the lack of consistency. The theoretical foundations of the Rasch model are much stronger than those of CTT, provided one endorses the revised definition of measurement.

Table 3
Fit statistics testing the experimental conditions based on the Rasch model.

Analysis	Condition	Overall fit	PSI	Conclusion
1	Entire sample	χ^2 (81) = 243.22 p < 0.0000001	0.75	Severe lack of fit implies heterogeneity across conditions.
2	Response scale A≫D	χ^2 (81) = 156.03 p = 0.000001	0.74	In both groups the overall fit is insufficient.
3	Response scale D≫A	χ^2 (81) = 165.75 p < 0.0000001	0.75	
4	Instruction: spontaneous (I-S)	χ^2 (81) = 156.37 p = 0.0000001	0.75	The well-considered instruction provides the best fitting response data, while the spontaneous instruction yields the worst fitting data. However, in all groups the overall fit is insufficient.
5	Instruction: well-considered (I-C)	χ^2 (81) = 125.48 p = 0.001	0.76	
6	Instruction: control group (I-CTRL)	χ^2 (81) = 129.83 p = 0.0004	0.74	
7	I-S A≫D	χ^2 (81) = 121.22 p = 0.003	0.75	Within each instruction condition, the A≫D scale works better, as expected. However, in the spontaneous condition, the difference is negligible.
8	I-S D≫A	χ^2 (81) = 123.10 p = 0.002	0.75	Within either response scale condition, the well-considered condition yields the best fitting data, while the spontaneous conditions performs worst.
9	I-C A≫D	χ^2 (81) = 81.66 p = 0.46	0.75	
10	I-C D≫A	χ^2 (81) = 106.10 p = 0.03	0.76	
11	I-CTRL A≫D	χ^2 (81) = 90.29 p = 0.23	0.75	
12	I-CTRL D≫A	χ^2 (81) = 106.47 p = 0.03	0.76	
13	Fast A≫D	χ^2 (81) = 111.70 p = 0.01	0.78	The A≫D version works better than the D≫A version for respondents who answer fast or at average speed (medium).
14	Fast D≫A	χ^2 (81) = 113.71 p = 0.01	0.77	
15	Medium A≫D	χ^2 (81) = 84.75 p = 0.37	0.70	The best fit occurs for respondents whose answer speed is average, on the A≫D format.
16	Medium D≫A	χ^2 (81) = 116.36 p = 0.01	0.75	
17	Slow A≫D	χ^2 (81) = 95.42 p = 0.13	0.70	
18	Slow D≫A	χ^2 (81) = 92.42 p = 0.18	0.74	

Table 4
Fit statistics of analyses testing the experimental conditions based on CFA (Mplus).

Analysis	Condition	χ^2 (df = 27)	RSMEA	Sample-size adjusted BIC
1	Entire sample	558.79	0.109	49,420.5
2	Response scale A≫D	326.56	0.116	24,621.7
3	Response scale D≫A	260.78	0.103	24,793.1
4	Instruction: spontaneous (I-S)	168.55	0.098	16,375.7
5	Instruction: well-considered (I-C)	253.35	0.118	16,688.3
6	Instruction: control group (I-CTRL)	220.67	0.114	16,427.2
7	I-S A≫D	97.07	0.098	7,981.4
8	I-S D≫A	103.52	0.102	8,329.3
9	I-C A≫D	172.46	0.139	8,502.9
10	I-C D≫A	105.32	0.103	8,187.0
11	I-CTRL A≫D	106.94	0.104	8,151.0
12	I-CTRL D≫A	135.03	0.121	8,304.6

An inspection of the data shows that response patterns based on spontaneous responses differ from those of well-considered responses in two ways. First, more respondents simply replicate their choice of a manifest response category across most items, implying artificial consistency. Second, more respondents disregard intermediate categories and instead resort to extreme responses. In other words, spontaneous responses promote particular response styles. In both instances, actual responses deviate from expected ones under the Rasch model, contributing to the lack of fit. In the Rasch model, respondents should agree with items representing relatively less of the property being measured and disagree with those representing relatively more. Consistent responses to all items run contrary to this expectation. In factor analysis, consistent responses, as well as excessively frequent extreme responses, lead to higher inter-item correlations, resulting in apparently higher item reliability.

The reverse is true for well-considered responses. Since the respondents better account for the item content, and therefore its location, their response patterns correspond more closely to what the Rasch model expects. However, the enhanced conformity of the responses with the Rasch model comes at the expense of response consistency. Hence, item intercorrelations decrease and become inconsistent. Therefore CFA fit deteriorates. The same logic applies to responses based on the D≫A format compared to A≫D. Consequently, the better fit of the spontaneous response data and the D≫A responses in CFA seems to be a methodological artifact.

6. Conclusions, discussion and implications

6.1. A new paradigm of measurement

Measurement in marketing research uses a multitude of diverse approaches which claim to serve the same purpose, quantifying a latent variable. The theoretical review of these approaches in this study raises doubts about the predominant paradigm of CTT but also about most suggested alternatives. Generally speaking, diversity of methods is a great strength in science. However, quantification is a very rigorous concept. The plurality of measurement approaches in marketing research is only possible due to defining measurement as the *assignment* of numerals. This definition has very little, if anything, to do with the universal scientific concept of quantity. A definition of measurement unique to the social sciences, which obscures the essence of quantity, is neither necessary nor useful. The suggestion of a quantitative latent variable implies an ontological claim. Addressing this claim means providing evidence that the variable is real and adheres to the structure of quantity. Failing to do so is tantamount to speculation, which is harmful to science in the long run. When Stevens invented his definition of measurement by assignment, the social sciences lacked the methodological repertoire and possibly also the understanding of how to tackle measurement in the way the natural sciences do. Today, the social sciences have the necessary models at their disposal and only adherence to tradition prevents researchers from working towards measurement that is on a par with the natural sciences.

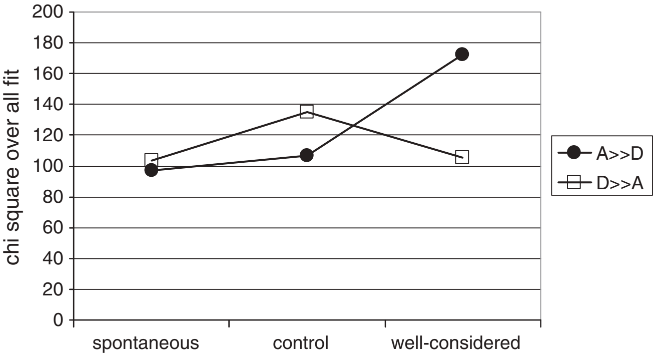

Fig. 4. Chi-square fit statistics for all six experimental conditions (CFA).

A revised definition of measurement based on revealing the structure of quantity clarifies what measurement really means. A special type of IRT, the Rasch model, does exactly that. Several published studies demonstrate the model's applicability to marketing. The present study adds to this body of evidence, although justification is not an empirical problem but requires theoretical considerations. The example presented reveals that the Rasch model and traditional CFA may lead to opposite conclusions, with those drawn from CFA being rather implausible. Nonetheless, many researchers have reservations about the use of the Rasch model. Most notably, many scholars still dispute the model's applicability for attitude scales. However, this objection is a misunderstanding and totally unfounded. The notion of an item representing more or less of a property applies to physical and attitudinal variables.

On the other hand, the Rasch model is no panacea to all measurement woes. For example, the sample invariance is not an easy solution to inappropriate sampling but a requirement the data must fulfill. Given evidence of invariance, researchers can exploit its power. Invariance allows for the comparison of measures based on different sets of items, thus contributing to a unification of multiple ways to measure essentially the same construct.

Measurement based on the Rasch model implies an unrivaled level of scientific justification, but measurement will not become easier. On the contrary, more deficiencies will become obvious and, in all likelihood, more attempts at measurement will prove unsuccessful. On the other hand, good scales based on CTT principles should withstand re-analysis by the Rasch model, for all intents and purposes. CTT is not wrong as such. It just fails to address the very problem of measurement. Under favorable circumstances, a CTT scale approximates measurement. A Rasch analysis can show whether this is indeed the case. At any rate, the Rasch model gives better directions in terms of how to improve an existing instrument.

Finally, the Rasch model promotes the development of more powerful substantial theories of constructs. Traditionally, content validity addresses facets of a construct. In contrast, the Rasch philosophy additionally asks for the characterization of different levels of the latent variable. The theory should allow for a prediction of the item location. The comparison of these theoretically expected and empirical item estimates considerably enhances the link between content and construct validity. Without the reference to content validity, the fit of the data to the Rasch model may just be a statistical routine. At present, most existing theories of constructs in marketing arguably lack the ability to predict the order of item locations. This shortcoming also illustrates the adverse effects of Stevens' unsuitable definition of measurement. Thus, researchers must advance the substantial construct theories and at least demonstrate the plausibility of the order of empirical item locations. Ideal definitions of constructs specify construct-specification equations (Stenner & Smith, 1982; Stenner, Smith, & Burdick, 1983). Such equations allow for concrete quantitative predictions of item locations. Currently, examples of such sophisticated measurement are scarce (e.g., the Lexile framework for reading ability, Lennon & Burdick, 2004); none seem to exist in the realm of attitude assessment. Whether measurement in marketing will ever reach this level remains to be seen.

6.2. Response scale direction

In general, the Rasch analysis favors the $A \gg D$ response scale format in the present study. This scale direction seems to facilitate the response process. However, respondents cope better with the $D \gg A$ format when they spend more time considering their responses. Instructions to provide well-considered responses are an extrinsic trigger in this context. Other circumstances might also prompt such behavior. High involvement, as a result of interest in the topic or when much is at stake for the respondent, is a possible intrinsic trigger of well-thought-out responses. The conclusion that $A \gg D$ is superior to $D \gg A$ represents only preliminary advice. Replications should shed more light on this issue. Specifically, further research should more closely investigate the functioning of differently directed response scales for reversed items. Finally, studies need to challenge the heuristic of spatial proximity.

6.3. Response instructions

When researchers request spontaneous responses, they often do so in order to counteract social desirability bias. Environmental attitudes constitute a construct which should be prone to this distortion. However, the Rasch analysis provides no indication that spontaneous responses are less socially desirable ones. In fact, the data based on spontaneous responses fit the model considerably worse. Consequently, asking for well-considered responses or refraining from mentioning response speed seems to be the better option subject to further empirical investigations.

In the present study, the respondents largely complied with instructions. However, the generalizability of this feature may be questionable. If respondents do not read instructions carefully, or are unwilling to comply, then the effects seen here might dissolve. In practice, the situation can be very complex. High involvement might interfere with a request for spontaneous responses. A long questionnaire might counteract a request for well-considered responses. In any case, Rasch analysis offers a powerful tool for putting all these conjectures to the test in the context of measurement.

References

Allan, S. J., & Hubbard, R. (1986). Regression equations for the latent roots of random data correlation matrices with unities on the diagonal. *Multivariate Behavioral Research*, *21*(3), 393–396.

Andrich, D. (1978a). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561–573.

Andrich, D. (1978b). Application of a psychometric rating model to ordered categories which are scored with successive integers. *Applied Psychological Measurement*, *2*(4), 581–594.

Andrich, D. (1982). An index of person separation in Latent Trait Theory, the traditional KR-20 Index, and the Guttman Scale Response Pattern. *Education Research and Perspectives*, *9*(1), 95–104.

Andrich, D. (1988). A general form of Rasch's extended logistic model for partial credit scoring. *Applied Measurement in Education*, *1*(4), 363–378.

Andrich, D. (1995a). Further remarks on non-dichotomization of graded responses. *Psychometrika*, *60*(1), 37–46.

Andrich, D. (1995b). Models for measurement, precision and the non-dichotomization of graded responses. *Psychometrika*, *60*(1), 7–26.

Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? *Medical Care*, *42*, 7–16.

Andrich, D., & Luo, G. (2003). Conditional pairwise estimation in the Rasch model for ordered response categories using principal components. *Journal of Applied Measurement*, *4*(3), 205–221.

Andrich, D., Sheridan, B. S., & Luo, G. (2010). *Rumm 2030: Rasch Unidimensional Measurement Models.* Western Australia: RUMM Laboratory Perth.

Balnaves, M., & Caputi, P. (2001). *Introduction to quantitative research methods, an investigative approach.* London: Sage.

Bechtel, G. G. (1985). Generalizing the Rasch model for consumer rating scales. *Marketing Science*, *4*(1), 66–77.

Bechtel, G. G., & Wiley, J. B. (1983). Probabilistic measurement of attributes: A logit analysis by generalized least squares. *Marketing Science*, *2*(4), 389–405.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord, & M. R. Novick (Eds.), *Statistical theories of mental test scores (Chs. 17–20)*. Reading, MA: Addison-Wesley.

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics.* Cambridge: Cambridge University Press.

Bradburn, N., Sudman, S., & Wansink, B. (2004). *Asking questions — The definitive guide to questionnaire design — For market research, political polls, and social and health questionnaires.* San Francisco, CA: Jossey-Bass.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. Bollen, & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). London & New Delhi: SAGE Publications.

Bruner, G. C., & Hensel, P. J. (1996). *Marketing scales handbook: A compilation of multi-item measures.* Chicago, IL: American Marketing Association.

Childers, T. L., & Heckler, S. E. (1985, September). Measurement of individual differences in visual versus verbal information processing. *Journal of Consumer Research*, *12*, 125–134.

Churchill, G. A. (1979, February). A paradigm for developing better measures of marketing constructs. *Journal of Marketing Research*, *XVI*, 64–73.

Creswell, J. W. (1994). *Research design: Qualitative and quantitative approaches.* London: Sage.

Dahlstrom, M., Nygaard, A., & Crosno, J. L. (2008). Strategic, metric, and methodological trends in marketing research and their implications for future theory and practice. *Journal of Marketing Theory and Practice*, *16*(2), 139–152.

Day, G. S., & Montgomery, D. B. (1999). Charting new directions for marketing. *Journal of Marketing*, *63*, 3–13.

Diamantopoulos, A. (2005). The C-OAR-SE procedure for scale development in marketing: A comment. *International Journal of Research in Marketing*, *22*(1), 1–9.

Diamantopoulos, A., & Winklhofer, H. M. (2001, May). Index construction with formative indicators: An alternative to scale development. *Journal of Marketing Research*, *38*, 269–277.

Dickson, J., & Albaum, G. (1975). Effects of polarity of semantic differential scales in consumer research. *Advances in Consumer Research*, *2*, 507–514.

Dunlap, R. E., & Van Liere, K. D. (1978). The “new environmental paradigm”: A proposed measuring instrument and preliminary results. *The Journal of Environmental Education*, *9*, 10–19.

Dunlap, R. E., Van Liere, K. D., Mertig, A. G., & Jones, R. E. (2000). Measuring endorsement of the New Ecological Paradigm: A revised NEP scale. *Journal of Social Issues*, *56*(3), 425–442.

Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods*, *5*(2), 155–174.

Ewing, M., Salzberger, T., & Sinkovics, R. (2005). An alternate approach to assessing cross-cultural measurement equivalence in advertising research. *Journal of Advertising*, *34*(1), 17–36.

Finn, A., & Kayande, U. (2005). How fine is C-OAR-SE? A generalizability theory perspective on Rossiter's procedure. *International Journal of Research in Marketing*, *22*(1), 11–21.

Friedman, H. H., Friedman, L. W., & Gluck, B. (1988). The effects of scale-checking styles on responses to a semantic differential scale. *Journal of the Marketing Research Society*, *30*(4), 477–481.

Friedman, H. H., Herskovitz, P. J., & Pollack, S. (1994). The biasing effects of scale-checking styles on response to a likert scale. *Proceedings of the American Statistical Association Annual Conference: Survey Research Methods* (pp. 477–481).

Ganglmair, A., & Lawson, R. (2003a). Advantages of Rasch modelling for the development of a scale to measure affective response to consumption. *European Advances in Consumer Research*, *6*, 162–168.

Ganglmair, A., & Lawson, R. (2003b). Measuring affective response to consumption using Rasch modelling. *Journal of Customer Satisfaction, Dissatisfaction and Complaining Behavior*, *16*, 198–210.

Glöckner, A., & Betsch, T. (2008). Multiple-reason decision making based on automatic processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(5), 1055–1075.

Griskevicius, V., Tybur, J. M., & Van den Bergh, B. (2010). Going green to be seen: Status, reputation, and conspicious conservation. *Journal of Personality and Social Psychology*, *98*(3), 392–404.

Hambleton, R. K., Swaminathan, H., & Rogers, J. H. (1991). *Fundamentals of item response theory.* Newbury Park: Sage Publications.

Hölder, O. (1901). Die Axiome der Quantität und die Lehre vom Mass. *Berichte über die Verhandlungen der Königlichen Sächsischen Gesellschaft der Wissenschaften zu Leipzig. Mathemathische-Physische Classe*, *35*. (pp. 1–64) Leipzig: Hirzel.

Holmes, C. (1974). A statistical evaluation of rating scales. *Journal of the Market Research Society*, *16*, 87–107.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*, 179–186.

Howell, R. D., Breivik, E., & Wilcox, J. B. (2007). Reconsidering formative measurement. *Psychological Methods*, *12*(2), 205–218.

Jacoby, J. (1978, April). Consumer research: A state of the art review. *Journal of Marketing*, *42*, 87–96.

Jarvis, C. B., MacKenzie, S. B., & Podsakoff, P. M. (2003, September). A critical review of construct indicators and measurement model misspecification in marketing and consumer research. *Journal of Consumer Research*, *30*, 199–218.

Jöreskog, K. G. (1971). Statistical analyses of sets of congeneric tests. *Psychometrika*, *36*, 109–133.

Jöreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association*, *70*(351), 631–639.

Karabatsos, G. (2001). The Rasch model, additive conjoint measurement, and new models of probabilistic measurement theory. *Journal of Applied Measurement*, *2*(4), 389–423.

Kilbourne, W. E., Beckmann, S. C., & Thelen, E. (2002). The role of the dominant social paradigm in environmental attitudes. A multinational examination. *Journal of Business Research*, *55*(3), 193–204.

Lee, N., & Hooley, G. (2005). The evolution of "classical mythology" within marketing measure development. *European Journal of Marketing*, *39*(3/4), 365–385.

Lennon, C., & Burdick, H. (2004). *The Lexile Framework as an approach for reading measurement and success.* Durham, NC: MetaMetrics, Inc.

Lievens, F., Sackett, P. R., & Buyse, T. (2009). The effects of response instructions on situational judgment test performance and validity in a high-stakes context. *Journal of Applied Psychology*, *94*(4), 1095–1101.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, *22*(140), 1–55.

Lord, F. M., & Novick, M. R. (Eds.). (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Luce, R. D., & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology*, *1*, 1–27.

MacKenzie, S. B. (2003). The dangers of poor construct conceptualization. *Journal of the Academy of Marketing Science*, *31*(3), 323–326.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174.

Mathews, C. O. (1929). The effect of printed response words on an interest questionnaire. *Journal of Educational Psychology*, *20*(2), 128–134.

McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & Lee Grubb, W., III (2007). Situational judgment tests, response instructions, and validity: A meta analysis. *Personnel Psychology*, *60*, 63–91.

Michell, J. (1986). Measurement scales and statistics: a clash of paradigms. *Psychological Bulletin*, *100*, 398–407.

Michell, J. (1990). *An introduction to the logic of psychological measurement.* Hillsdale, NJ: Erlbaum.

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology*, *88*, 355–383.

Michell, J. (1999). *Measurement in psychology — A critical history of a methodological concept.* Cambridge: Cambridge University Press.

Michell, J. (2003). Measurement: A beginner's guide. *Journal of Applied Measurement*, *4*(4), 298–308.

Michell, J., & Ernst, C. (1996). The axioms of quantity and the theory of measurement, part I, an english translation of Hölder (1901). *Journal of Mathematical Psychology*, *40*, 235–252.

Michell, J., & Ernst, C. (1997). The axioms of quantity and the theory of measurement, part II, an english translation of Hölder (1901). *Journal of Mathematical Psychology*, *41*, 345–356.

Molenaar, I. W. (1995). Estimation of item parameters. In G. H. Fischer, & I. W. Molenaar (Eds.), *Rasch models, foundations recent developments, and applications* (pp. 39–51). New York, NY: Springer.

Muthén, B., & Muthén, L. (2008). Mplus. http://www.statmodel.com/ (computer software. Los Angeles).

Nimkar, C. Y. (2010). Questionnaire design. http://1066www.scribd.com/doc/25419095/Questionnaire-Design (Retrieved June 10, 2010).

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw-Hill.

Ping, R. A., Jr. (2004). On assuring valid measurement for theoretical models using survey data. *Journal of Business Research*, *57*, 125–141.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests, copenhagen: danish institute for educational research (reprint 1980, expanded ed. with foreword and afterword by B. D. Wright).* Chicago: The University of Chicago Press.

Rasch, G. (1977). On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements. *Danish Yearbook of Philosophy*, *14*, 58–93.

Rossiter, J. R. (2002). The C-OAR-SE procedure for scale development in marketing. *International Journal of Research in Marketing*, *19*, 305–335.

Rossiter, J. R. (2010). *Measurement for the social sciences. The C-OAR-SE method and why it must replace psychometrics.* New York, NY: Springer.

Salzberger, T., & Sinkovics, R. (2006). Reconsidering the problem of data equivalence in international marketing research — Contrasting approaches based on CFA and the Rasch model for measurement. *International Marketing Review*, *23*(4), 390–417.

Sheluga, D., Jacoby, J., & Major, B. (1978). Whether to agree–disagree or disagree–agree: The effects of anchor order on item response. *Advances in Consumer Research*, *5*(1), 109–113.

Singh, J. (2004). Tackling measurement problems with item response theory: Principles, characteristics, and assessment, with an illustrative example. *Journal of Business Research*, *57*(2), 184–208.

Smith, E. V. (2002). Understanding Rasch measurement: detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement*, *3*(2), 205–230.

Stenner, A. J., Burdick, D. S., & Stone, M. H. (2008). Formative and reflective models: Can a Rasch analysis tell the difference? *Rasch Measurement Transactions*, *22*(1), 1152–1153.

Stenner, A. J., & Smith, M. (1982). Testing construct theories. *Journal of Perceptual and Motor Skills*, *55*, 415–426.

Stenner, A. J., Smith, M., & Burdick, D. S. (1983). Toward a theory of construct definition. *Journal of Education Measurement*, *20*(4), 305–316.

Stenner, A. J., Stone, M. H., & Burdick, D. S. (2009). Indexing vs. measuring. *Rasch Measurement Transactions*, *22*(4), 1176–1177.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*, 667–680.

Stevens, S. S. (1951). Mathematics, measurement, and psychophysics. In S. S. Stevens (Ed.), *Handbook of experimental psychology* (pp. 1–49). New York: Wiley.

Stevens, S. S. (1959). Measurement, psychophysics and utility. In C. W. Churchman, & P. Ratoosh (Eds.), *Measurement: Definitions and theories* (pp. 18–63). New York, NY: John Wiley (Ch. 2).

Stewart, D. W. (1981). The application and misapplication of factor analysis in marketing research. *Journal of Marketing Research*, *18*(1), 51–62.

Thurstone, L. L. (1928). Attitudes can be measured. *The American Journal of Sociology*, *33*, 529–554.

Tourangeau, R., Couper, M. P., & Conrad, F. (2004). Spacing, position, and order, interpretive heuristics for visual features of survey questions. *Public Opinion Quarterly*, *68*(3), 368–393.

Tourangeau, R., & Rasinski, K. A. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin*, *103*(3), 299–314.

Venkatraman, M. P. (1991, Spring). The impact of innovativeness and innovation type on adpotion. *Journal of Retailing*, *67*, 51–67.

Venkatraman, M. P., & Price, L. L. (1990, June). Differentiating between cognitive and sensory innovativeness: Concepts, measurement, and implications. *Journal of Business Research*, *20*, 293–315.

Watkins, M. M. (2000). *Monte Carlo PCA for parallel analysis [computer software].* State College, PA: Ed & Psych Associates (The program can be downloaded from). http://freewareapp.com/monte-carlo-pca-for-parallel-analysis_download (last accessed June 30, 2010).

Weng, L. (2004). Impact of the number of response categories and anchor labels on coefficient alpha and test–retest reliability. *Educational and Psychological Measurement*, *64*(6), 956–972.

Weng, L., & Cheng, C. -P. (2000). Effects of response order on Likert-type scales. *Educational and Psychological Measurement*, *60*(6), 908–924.

Wenke, D., & Frensch, P. A. (2005). The influence of task instruction on action coding: Constraint setting or direct coding? *Journal of Experimental Psychology. Human Perception and Performance*, *31*, 803–819.

Wilcox, J. B., Howell, R. D., & Breivik, E. (2008). Questions about formative measurement. *Journal of Business Research*, *61*, 1219–1228.

Wilson, M. (2005). *Constructing measures, An item response modeling approach.* Mahwah, NJ: Lawrence Erlbaum Associates.

World Health Organization (2009). Perception survey for health-careworkers. http://www.who.int/entity/gpsc/5may/Perception_Survey_for_Senior_Managers.doc (Retrieved June 10, 2010).

Zwindermann, A. H. (1995). Pairwise parameter estimation in Rasch models. *Applied Psychological Measurement*, *19*, 369–375.