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