Thomas Salzberger and Monika Koller

The direction of the response scale matters - accounting for the unit of measurement

Article (Published)
(Refereed)

European Journal of Marketing

The direction of the response scale matters - accounting for the unit of measurement

Thomas Salzberger, Monika Koller,

The direction of the response scale matters – accounting for the unit of measurement

Thomas Salzberger and Monika Koller
Department of Marketing, WU Vienna, Vienna, Austria

Received 24 August 2017
Revised 2 March 2018
13 July 2018
23 July 2018
Accepted 27 July 2018

Abstract

Purpose – Psychometric analyses of self-administered questionnaire data tend to focus on items and instruments as a whole. The purpose of this paper is to investigate the functioning of the response scale and its impact on measurement precision. In terms of the response scale direction, existing evidence is mixed and inconclusive.

Design/methodology/approach – Three experiments are conducted to examine the functioning of response scales of different direction, ranging from agree to disagree versus from disagree to agree. The response scale direction effect is exemplified by two different latent constructs by applying the Rasch model for measurement.

Findings – The agree-to-disagree format generally performs better than the disagree-to-agree variant with spatial proximity between the statement and the agree-pole of the scale appearing to drive the effect. The difference is essentially related to the unit of measurement.

Research limitations/implications – A careful investigation of the functioning of the response scale should be part of every psychometric assessment. The framework of Rasch measurement theory offers unique opportunities in this regard.

Practical implications – Besides content, validity and reliability, academics and practitioners utilising published measurement instruments are advised to consider any evidence on the response scale functioning that is available.

Originality/value – The study exemplifies the application of the Rasch model to assess measurement precision as a function of the design of the response scale. The methodology raises the awareness for the unit of measurement, which typically remains hidden.

Keywords Surveys, Eye tracking, Measurement precision, Unit of measurement

Paper type Research paper

Introduction

Measurement plays a vital role in the success story of the natural sciences with the social sciences attempting to follow this role model. In the natural sciences, both the importance of a common unit of measurement, such as 1 m, or 1 kg (Bordé, 2005), and an estimate of

The authors are grateful to the editor-in-chief, Greg W. Marshall, the associate editor, Kevin Voss, and four anonymous reviewers for their helpful guidance and suggestions. In addition, the authors thank Christoph Himmer for assistance in data collection, and Steve Humphrey and David Andrich (University of Western Australia) for their valuable input and guidance.

European Journal of Marketing
Emerald Publishing Limited
0309-0566
DOI 10.1108/EJM-08-2017-0539

measurement precision (BIPM *et al.*, 2008) are universally recognised. In social measurement, awareness of these constituents of measurement has risen only recently (Pendrill and Fisher, 2013; Pendrill *et al.*, 2017). Currently, social measurement, mostly, lacks meaningful estimates of precision. In large part, this appears to be due to limitations in the methods used.

Factor analysis yields respondent measures that are percentiles of a distribution assumed to be normal. Hence the implicit unit of measurement is intrinsically tied to a given population and context of measurement seriously hampering the generalisability of measures and the discovery of universal laws in the social sciences.

Reliability, commonly considered the hallmark of an instrument's precision, suffers from similar shortcomings that apply to the measurement unit (Voss *et al.*, 2000), as it confounds properties of the respondents (their true variance) and measurement precision (error variance). The standard error of measurement (SEM) derived from reliability (Rel), SEM = standard deviation $\sqrt{1 - Rel}$ (Traub, 1994), also depends on sample characteristics. What is more, it is the same regardless of the magnitude of the measure. In practice, every instrument is more precise in a range to which it is targeted and becomes less precise when approaching its boundaries. Paradoxically, consistency is higher at the extremes demonstrating that it does not necessarily indicate precision. Therefore it is important to investigate the role of the unit of measurement, its relation to precision, and how it can be assessed.

Measurement precision

While the required precision depends on the purpose of measurement, as a matter of principle, measurement conditions should be designed with the objective of maximising precision. Ideally, the investigation of those conditions is carried out in an experimental setting prior to large scale data collection as part of a comprehensive scale development project or a methodological study. In terms of the psychometric analysis, modern test theory (Andrich, 2002, 2011) lends itself as an alternative approach as it allows for disentangling the measurement unit from distributional properties of the sample.

Precision is intrinsically tied to the unit of measurement (Humphry, 2005; Humphry and Andrich, 2008). For example, measures of length estimated in millimetres are more precise (and, incidentally, less consistent on replication) than measures estimated in centimetres. Thus, the meaningful interpretation of the standard error of measurement hinges on the unit it is expressed in. From this perspective, stating a range of uncertainty only makes sense once validity has been sufficiently supported. By contrast, in the traditional paradigm, the assessment of reliability usually precedes the investigation of validity (Voss *et al.*, 2000). However, factors impacting measurement precision, such as the tendency to provide socially desirable responses (Steenkamp *et al.*, 2010) or response sets (Greenleaf, 1992; Baumgartner and Steenkamp, 2001), may also compromise validity. Every component involved in measurement, e.g. the instrument, its layout, its administration, the characteristics of the respondents, the context of data collection, may influence precision. In the following, the effect of the direction of the response scale will be examined as one factor impacting measurement precision.

The response scale

The design of the response scale and its relationship to precision has long been of interest to marketing scholars. Generally, more response categories imply higher precision (Lozano *et al.*, 2008), provided respondents do not get overburdened resulting in dropouts (Galesic, 2006) or response sets disregarding some of the response options. The verbalisation of categories (Menold and Tausch, 2016) or adding numeric scores to the response categories, the number of options (Weijters *et al.*, 2010), or the visual representation (Parasuraman *et al.*, 1998) may also

have an impact. The direction of the response scale has attracted comparatively less interest. In the context of agreement versus disagreement, the response scale may either run from agree to disagree, hereinafter denoted agree-to-disagree, or, conversely, start with disagree progressing to agree, referred to as disagree-to-agree (Table I).

When designing the response scale, researchers typically rely on previously used formats. Even methodological papers proposing procedures of scale development barely deal with the response scale and, specifically, its direction (MacKenzie *et al.*, 2011), which, by and large, appears to be extraneous to precision and validity. This impression is reinforced by the fact that most authors of empirical studies fail to report explicitly which scale direction they used. The scarcity of information makes it difficult to get an idea of which response scale direction is more popular. As trivial and inconsequential a simple lateral transposition of the response options may seem, it may, in principle, affect measurement in various ways. If such effects go unnoticed, suboptimal response scale formats might be used or problems of non-comparability occur.

The direction of the response scale has been revisited intermittently in the past century (Mathews, 1929; Sheluga *et al.*, 1978). Recently, Yan and Keusch (2015) summarised pertinent studies in this regard. Their conclusion reveals that findings are mixed and inconclusive. The unsatisfactory state of insight into the problem is further documented by the fact that the studies undertaken are typically descriptive in nature offering no explanatory mechanism responsible for occasionally observed, but inconsistent, effects. Another major limitation is the focus on comparisons of means and variances of ordinal raw scores. So far, no light has been shed on the impact of the scale direction on the unit of measurement and, by implication, on measurement precision at the level of the latent variable.

Hence, the present paper investigates potential effects of the direction of the response scale on the unit of measurement in a series of three experiments. The first two studies involve psychometric analyses based on the Rasch model (RM) for measurement (Rasch, 1960; Andrich, 1988). The third study investigates eye movements recorded during the completion of a computer-administered questionnaire.

In a between-subjects design, Experiment 1 investigates whether the response scale direction has an effect on the underlying metric using the CETSCALE (Shimp and Sharma, 1987) in a paper-and-pencil administration as an exemplar.

Experiment 2 extends the scope of examination to data collected online. Furthermore, the experiment investigates a suggested mechanism responsible for the potential occurrence of a response scale direction effect.

In Experiment 3, physiological data are used to examine the reading patters of respondents when using rating scales with alternate scale directions. Specifically, it is tested whether gaze motions differ depending on the response scale format presented in a between subject design.

Theoretical framework

The attributes of the response scale

Many aspects of the design of the response scale have been investigated in previous research. The appropriate number of response categories (Weng, 2004; Cox, 1980), odd

(a)	*Item statement*	agree disagree ☐ ☐ ☐ ☐ ☐	Agree-to-disagree format
(b)	*Item statement*	disagree agree ☐ ☐ ☐ ☐ ☐	Disagree-to-agree format

Table I.
Response scale formats of different direction

versus even-numbered categories (Wong *et al.*, 1993), the spacing between categories (Rea and Parker, 2014), general formatting issues (Fanning, 2005) and the verbal or numerical labelling of response options (Rammstedt and Krebs, 2007; Weijters *et al.*, 2010), to name the most important issues, have been considered.

In terms of the direction of the response scale, existing studies focus on the potential for additive biases while occasionally also scrutinising distributional properties and reliability. The evidence in the literature is mixed, though. Dickson and Albaum (1975) found no mean differences with semantic differential scales, while Mathews (1929) found a bias towards the left. Likewise, Sheluga *et al.* (1978) and Friedman *et al.* (1988, 1994) identified stronger agreement for the agree-to-disagree format, when items were worded favourably, but no differences in terms of variances and reliability. Rammstedt and Krebs (2007) identified no differences in means and variances provided numerical labels matched the verbal labels (i.e. a low number implies disagreement and a high number agreement). Salzberger and Koller (2013) found an interaction between response instructions ranging from well-considered to spontaneous and the response scale direction. Liu and Keusch (2017) reported an effect on response styles such as an acquiescent bias in web surveys but not in face-to-face settings. Bradburn *et al.* (2004, p. 161) concluded that there is "no good evidence that one form is universally better than another", as multiple factors apparently interact. Further progress seems to require more advanced diagnostic tools, a clearer conceptualisation of possible effects, and a better theoretical framework of possible mechanisms.

Psychometric analysis of the response scale

Classical test theory (CTT; Lord and Novick, 1968) offers limited possibilities of investigating the effects on precision for two reasons. First, CTT statistics such as correlations, variances, and reliability are sample-dependent. Second, item raw scores are treated as a priori meaningful statistics and response categories are not parameterised (Andrich, 2011).

The application of the RM (Rasch, 1960; Andrich, 1988) overcomes these limitations of CTT. Among the broader group of Item Response Theory models (Embretson and Reise, 2013; Raykov and Calantone, 2014), the RM is unique with respect to the separation of item and respondent characteristics (Fischer, 1995) allowing for item parameter estimates to be statistically independent of the distribution of respondents. The popularity of the RM in marketing has increased sharply in recent years (Salzberger, 2009; Ramón Oreja-Rodríguez and Yanes-Estévez, 2010; Ganglmair-Wooliscroft and Wooliscroft, 2013; Salzberger and Koller, 2013; Salzberger *et al.*, 2014; Sweeney *et al.*, 2015; Ganglmair-Wooliscroft and Wooliscroft, 2016).

The dichotomous RM [equation (1)] links the observed response a_{vi} to an item i by a respondent v by a logistic function of person (β_v) and item parameters (δ_i), which are estimated from the data and expressed in the same metric. The person parameter β_v represents the measure for an individual respondent corresponding to the factor score in factor analysis. The item parameter δ_i represents the location of the item on the latent continuum. Its closest counterpart in CTT is the item mean, which, however, evidently depends on the sample. The larger β_v is and/or the smaller δ_i is, the more likely the respondent agrees with the item. The parameters β_v and δ_i are expressed in logits, or log-odds (Wright and Mok, 2000). The RM requires discrimination to be equal across items ensuring parameter separation (Fischer, 1995), which implies specific objectivity (Rasch, 1977) as the defining characteristic of any RM.

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

Polytomous response data require a parametrisation of the response categories by means of threshold parameters τ marking the boundaries between adjacent categories. In the RM response categories do not represent points on the latent continuum but cover a specific range, where they are expected to be the most likely response. In the binary case, the item parameter δ_i can also be seen as the threshold between a positive (e.g. agree) and a negative response (e.g. disagree). Hence, a seven-category response scale requires six τ parameters. In the context of this study, the threshold parameters are of particular importance as they reveal how the response categories actually work (Andrich, 2011).

Equation (2) shows the probability of choosing a particular response category according to the polytomous RM for ordered categories (Andrich, 1988, p. 366). The denominator γ is the sum of all numerators for all response categories. The parameter δ_i is the mean of all thresholds.

$$P\left(a_{vi} = \mathrm{x} \mid \beta_v, \tau_{ij}, \mathrm{j} = 1 \ldots \mathrm{m}, 0 < \mathrm{x} \leq \mathrm{m}\right) = \frac{e^{\left(\Sigma_{j=1}^{x} - \tau_{ij}\right) + \mathrm{x} \cdot (\beta_v - \delta_i)}}{\gamma} \quad (2)$$

with,

$$\gamma = 1 + \Sigma_{k=1}^{m} e^{\left(\Sigma_{j=1}^{x} - \tau_{ij}\right) + \mathrm{k} \cdot (\beta_v - \delta_i)} \quad (3)$$

$$\delta_i = \frac{\sum_{j=1}^{x} \tau_{ij}}{m} \quad (4)$$

$$P(a_{vi} = 0) = \frac{1}{\gamma} \quad (5)$$

As parameter estimates are only meaningful if the data fit the model, it has to be checked whether the assumptions implied by the RM actually hold true for the data. One fundamental fit statistic, which is approximately chi-square distributed, compares expected item scores with actually observed item scores (Andrich, 1978) in groups of respondents formed according to their location of the latent continuum. Summed across all items, it yields an overall fit statistic at the scale level. Other requirements, such as local independence, unidimensionality and lack of differential item functioning (no item bias), have to be addressed, too (Ewing *et al.*, 2005, and Salzberger, 2009, for details). In the empirical examples presented, results of tests of fit are only reported if problems were encountered.

Potential effects of the response scale on data quality

Three effects of the response scale can be distinguished. First, the psychometric properties of the items may be compromised to a degree that validity becomes questionable. Second, there may be an additive bias associated with one version relative to the other. Third, there may be a difference in the unit of measurement and precision.

The first effect can be investigated by the analysis of fit of the data to the RM. The second effect can be addressed by a mean comparison of person measures from two groups that are not supposed to differ, as is the case in a randomised experiment. The third effect is related to the degree to which respondents discriminate between response categories and different items. The latter determines the precision of measurement. The sharper the discrimination between two items i and j as evidenced by the number of people agreeing with item i but not with item j given that they agree with one and only one of the two items, the further apart δ_i and δ_j will be (Humphry, 2005; Humphry and Andrich, 2008). The same logic applies to distances between thresholds. This is directly analogous to two markings on a ruler being 1 cm apart. When measuring in centimetres, there is only one unit between the markings, whereas the more precise measurement in millimetres implies ten units between the same markings. Obviously, a measurement uncertainty of $\pm$ 5 units in millimetres implies more precise measurement than an uncertainty of $\pm$ 1 unit in centimetres.

A difference in the unit may imply a spurious mean effect depending on the locations of persons relative to the locations of items. When an additive bias and a change in the unit occur simultaneously, the size and the direction of the mean difference also depend on the relative locations of the respondents and the items.

The crux is that currently in the social sciences no awareness exists of the size of the (implicit) unit, let alone how one unit could be converted into another. In the natural sciences, measurement units are self-evident most notably because they are explicit and tangible.

Discrimination between response categories

When respondents discriminate differently between response categories depending on the response scale direction, the unit of measurement will be different. This kind of bias will be referred to as a multiplicative bias (Figure 1).

It is the fact that discrimination is set to unity in the RM that explains why precision works out this way. If respondents discriminate more sharply between response categories, the distances between thresholds get bigger and each category is represented by a wider range of the latent variable allowing for a higher granularity of measurement. The change in the metric also results in a wider spread of person measures resulting in better person separation.

Interpretive heuristics used by the respondents

Social measurement requires the interpretation of the item and the response scale by the respondent. Hence, not only the objective properties matter but also interpretive heuristics on the part of the respondent. The heuristic "near means close" (Tourangeau *et al.*, 2004, p. 370) based on concepts of Gestalt theory appears particularly promising when explaining the underlying psychological mechanism of a possible response scale direction effect. It

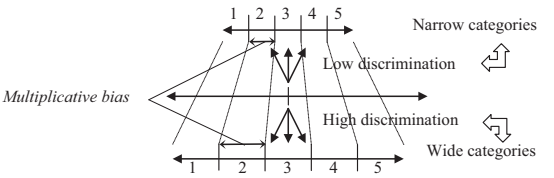

Figure 1.
Relationship of discrimination between response categories and distances between thresholds implying a difference in the unit of measurement

suggests that respondents interpret stimuli which are presented near to one another as more closely related compared to stimuli being spatially apart. While Tourangeau *et al.* (2004) investigated the spatial arrangement of items, it can be argued that this heuristic may also apply to the spatial relationship between the item and the associated response scale. While agreeing to an item implies 'feeling close to the item', rejection of the statement means psychological distance between the respondent and the item. Consequently, spatial proximity between the statement and the agree-pole of a Likert-type response scale of the agree-to-disagree format resonates with a favourable mental position *vis-à-vis* the item, while spatial distance between the item and the disagree pole accommodates a lack of congruence between the respondent and the item.

Running contrary to the near-means-close heuristic, the disagree-to-agree response scale could adversely affect the response process and the discrimination between the response categories by the respondents. The spatial proximity heuristic and its consequences should only apply when the scale is presented to the right of the item, though. There should be no difference when the response categories are positioned below the item, as all response options are then equidistant to the statement. In Experiment 2, this proposition is explicitly tested.

In terms of an additive bias, two proposed mechanisms seem to be particularly relevant (Yan and Keusch, 2015). A primacy effect suggests that respondents read response options sequentially from left to right and choose the first option that is sufficiently close to their true level of agreement. Such response behaviour has been referred to as satisficing (Krosnick, 1999), resulting in drop-out, straightlining, skipping responses, or resorting to a *don't know* option if offered.

Alternatively, the anchoring-and-adjustment heuristic (Tversky and Kahneman, 1974) proposes that respondents could perceive the first option to the left as an anchor and adjust their response in relation to that anchor. As a result, responses are expected to be closer to the anchor than they should be.

Both the primacy effect and the anchoring-and-adjustment heuristic should both be present irrespective of the positioning of the response scale to the right of the item or beneath. Both mechanisms come with theoretical limitations, though. In case of the primacy effect, it is plausible that respondents *read* the scale from left to right. Whether this also means that they *consider* the options in that order when actually responding is questionable. In fact, the RM assumes a response process where the probability of each response option depends on all other options, as the denominator in equation (2) includes all thresholds. Thus, the primacy effect should also result in item misfit to the RM.

The anchoring-and-adjustment heuristic has been proposed, and repeatedly confirmed, for numerical estimates, particularly under the condition of uncertainty. It is questionable whether selecting a response category that does not come with a numerical label attached to it, really matches these conditions. The assumption of uncertainty appears questionable as one would hope that respondents do know their true stance with respect to the item. Thus, primacy due to satisficing appears to be the more plausible mechanism in the present context, but it should also lead to item misfit and not just a mean shift.

Experiment 1

In Experiment 1, the potential scale direction effect is examined with the response scale being presented to the right of the item. Given the inconclusive evidence in the literature, the first experiment, which is based on a paper-and-pencil administration, is an exploratory pilot study with respect to the presence of additive and/or multiplicative bias.

A difference in the unit of measurement can be inferred provided three conditions are fulfilled (Humphry, 2005). First, the standard deviations of the item location and item threshold estimates differ between the agree-to-disagree scale and the disagree-to-agree scale. Second, one set of estimates can be transformed into the other by a multiplicative constant (the estimates from one scale are shrunk uniformly compared to the other). Third, the standard deviations of the person estimates differ by a multiplicative constant in the same order of magnitude as the standard deviations of the item locations.

Method

The CETSCALE (Shimp and Sharma, 1987, for item wording), a 17-item instrument measuring consumer ethnocentric tendencies, is used in Experiment 1. In the present study a published translation into the native language of the study participants has been used (Sinkovics, 1999). Any reference to America was replaced by the name of the country of the study participants. The instrument comes with a seven-point Likert-type response scale as suggested by its developers. The CETSCALE has proven to be quite robust and valid under many circumstances (Jiménez-Guerrero *et al.*, 2014). It has been used repeatedly as a showcase for methodological investigations in marketing (Clarke, 2001; Baumgartner and Steenkamp, 2001). Even though the CETSCALE has been developed according to CTT procedures, the application of the RM is meaningful for three reasons. First, the qualitative underpinning of the scale's construction should result in a sufficient number of items satisfying the RM. Second, if one adheres to the properties of the RM as being essential for social measurement, relying on CTT simply because a scale has historically been based on CTT is not particularly conducive. The RM may test to what extent the instrument fulfils these requirements. Third, the RM has been successfully applied to the CETSCALE in the past (Salzberger *et al.*, 1997).

Two different versions of the response scale were administered (i.e. agree-to-disagree, n = 146, versus disagree-to-agree, n = 149, endpoints anchored as *fully agree* and *fully disagree*, respectively) to a sample of first-year students at a Business School in Vienna, Austria. A between-subjects design with random assignment was used to avoid dependency between repeated measurements. This ensured that the two samples were stochastically equivalent, and that any effect could be attributed to the scale direction. The data were analysed by the RM for polytomous data (rating scale model) using RUMM 2030 (Andrich *et al.*, 2009-2012).

Results

Initially, data from the two conditions were pooled. Four items had to be deleted at that stage because of consistently showing strong misfit. With 13 items remaining, certainly a broad-enough basis of further analysis was provided.

Nonetheless, the data exhibited poor overall fit to the model (χ^2 = 103.24, df = 52, $p <$ 0.0001) suggesting further problems prevailing in the data. As misfit was one possible outcome of the response scale direction, the respective data were split. Model fit improved substantially for the agree-to-disagree version (χ^2 = 33.28, df = 26, p = 0.15), while fit remained poor for the disagree-to-agree format (χ^2 = 84.20, df = 26, $p <$ 0.0001) demonstrating that the good fit for agree-to-disagree data cannot be attributed to the smaller sample size alone (see Table II for a summary of the findings).

This finding suggests that the agree-to-disagree version is the more appropriate response format. While no significant mean difference was observed in the pooled data (−0.65 versus −0.60, p = 0.71), there was a difference in the standard deviations of respondent measures based on different response scale directions standing in a ratio of 1:1.32 (1.09:1.43). The standard deviations of the item location estimates for one version compared to the other

Response scale Criterion	Condition 1: Agree-To-Disagree	Condition 2: Disagree-To-Agree
Overall fit	$\chi^2 = 33.28$	$\chi^2 = 84.20$
(df)	(26), $p = 0.15$	(26), $p < 0.0001$
Mean of item fit χ^2 (df)	2.56 (2)	6.48 (2)
Standard deviation of item locations	0.57	0.43
Standard deviation of person locations	1.43	1.09
Standard deviation of person locations/ standard deviation of item locations	2.51	2.53
Relative unit (standard deviation of item locations divided by standard deviation of item locations in reference group *Condition 1)*	1.00	0.75
Mean of person locations	−0.71	−0.57
Adjusted mean of person locations (mean divided by relative unit)	−0.71	−0.76
Regression item location estimates on Condition 1	–	b = 0.74, $r^2 = 0.94$
Items retained	1, 3, 6*, 7*, 9, 10, 11*, 12, 13*, 14, 15, 16*, 17* (Shimp and Sharma, 1987, p.283); * indicates item is part of the 10-item version	

Table II.
Key results of the analyses of the CETSCALE in Experiment 1: effect of the response scale direction on fit, the unit of measurement and precision (paper-and-pencil)

stood in a very similar ratio of 1:1.33 (0.43:0.57) suggesting a difference in the unit of measurement. To confirm this finding, two plots of the actual item location estimates and the threshold estimates, respectively, from one version against those from the other were created. If the estimates approximate a straight line at a 45° angle, than they are equal and share the same unit. In Experiment 1, however, they approach a straight line at a different angle. Thus, one scale is shrunk compared to the other and, consequently, the unit of measurement is different (Figure 2).

To explore whether the phenomenon of poorer precision in case of the disagree-to-agree format can be subsumed under the concept of satisficing (Krosnick, 1999), the proportion of missing values and straightlining were investigated. Missing values were extremely rare under both conditions with only one participant featuring one missing value in case of agree-to-disagree and seven respondents with just one missing value in case of disagree-to-agree. Straightlining was observed for ten respondents (6.8 per cent) in case of agree-to-

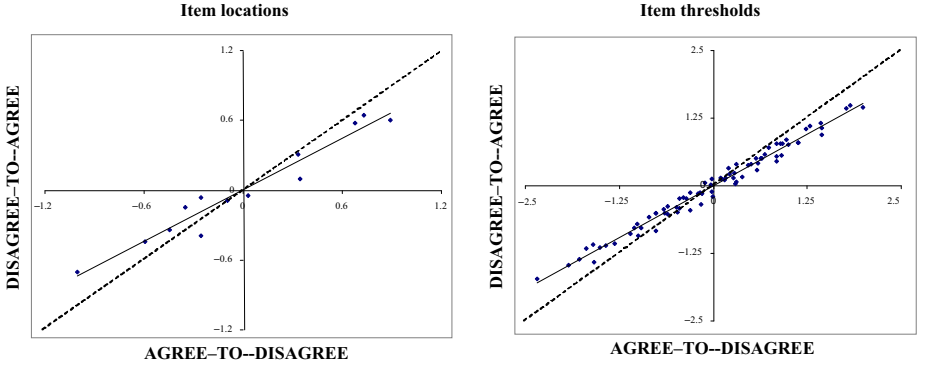

Figure 2.
Estimates of item locations and thresholds comparing agree-to-disagree and disagree-to-agree

disagree and five respondents (3.4 per cent) in case of disagree-to-agree. The difference in these proportions is statistically insignificant ($p = 0.18$). Thus, the observed impact on the quality of the data does not seem to be due to satisficing.

Discussion

The superiority of the agree-to-disagree format suggests that the direction of the response scale does matter in terms of precision of measurement. But do the data support the spatial proximity heuristic of near-means-close? Alternatively, it could be that respondents discriminate more sharply when their favourite categories are presented first, i.e. to the left. It might be more demanding, when the respondent, after reading the item, has to switch to the opposite extreme first and then select a particular category. This post-hoc hypothesis follows the argument of a primacy effect as a consequence of satisficing. It was tested by analysing sub-samples of respondents who generally agreed (high score group) versus those who predominantly disagreed (low score group). Both subsamples discriminated more sharply when using the agree-to-disagree format rather than the disagree-to-agree version providing no evidence of a primacy effect. The findings are consistent with the spatial proximity heuristic, though.

Experiment 2

Experiment 2 investigates whether the effect found in Experiment 1 can be replicated for online data as the more popular approach today. When using different data collection methods, for example to better reach different segments of the population (Dolnicar *et al.*, 2009), care should be taken when pooling the data (Deutskens *et al.*, 2006; Mead and Drasgow, 1993). In addition, more light will be shed on the mechanism that could be responsible for the difference between agree-to-disagree and disagree-to-agree.

Experiment 2 extended the investigation of the possible effect by also considering the response scale positioned below the item rather than to the right, as well as a third condition for the scale direction with agree-to-disagree and disagree-to-agree presented randomly for each item. Experiment 2 consisted of two parts. In Part 1, the effect of the response scale direction (agree-to-disagree versus disagree-to-agree versus random) was investigated. In Part 2, a 2×2 factorial design was implemented using a different instrument. Factor 1 featured the response scale direction (agree-to-disagree versus disagree-to-agree), while Factor 2 accounted for the placement of the response scale (to the right of the item versus below). Based on the spatial proximity heuristic, there should be no effect of the response scale direction when it is placed beneath the item.

Method

Data were collected online from a total of 1,647 respondents, mostly students from a European Business School ensuring the same population as in Experiment 1. The sample used was relatively homogeneous in terms of age (mean 25 years, 97 per cent are under 41). As in Experiment 1, the survey included the original 17-item version of the CETSCALE in Part 1. In Part 2, eight items measuring affective concern (affective concern, AFC, is a latent construct within the context of environmental concern of consumers, as proposed by Salzberger, 2007; see Appendix for item wording; the scale has been developed in the native language of the study participants) were administered using the same seven-category response scale. The CETSCALE items were presented in three different conditions with respondents being randomly assigned. An agree-to-disagree response scale was presented in Condition 1, respondents in Condition 2 were offered a disagree-to-agree scale, while in Condition 3 respondents got the agree-to-disagree scale or the disagree-to-agree scale for

each item at random. Changing the direction of the response scale after each item at random in Group 3 is expected to trigger as much confusion as can possibly be caused by the response format alone.

The AFC-items were presented in four different conditions. In Conditions 1 and 2, the response scale was presented to the right of the statement with the agree-to-disagree format used in Condition 1 and the disagree-to-agree format in Condition 2. In Conditions 3 and 4, the response scale was placed below the item. Again, the direction was varied with agree-to-disagree used in Condition 3 and disagree-to-agree in Condition 4.

Based on the spatial proximity hypothesis, the agree-to-disagree scales were expected to be superior to the disagree-to-agree version (better fit, higher precision) for the CETSCALE data and the AFC data when presented to the right of the items. The random variation of the response scale direction was expected to result in strong misfit due to the confusion it triggered. No difference in precision and fit between the agree-to-disagree and the disagree-to-agree formats was expected when the response scale was placed below the item.

All data analyses were again carried out based on the RM for polytomous data using RUMM 2030 (Andrich *et al.*, 2009-2012).

Results

Part 1: Analysis of the CETSCALE. At first, eight items were omitted from the analysis because of local dependence (LD; Marais and Andrich, 2008; Marais, 2013) as evidenced by correlations of item residuals (Yen, 1984; Marais and Andrich, 2008). LD implies redundancy, which was due to the high content similarity of some items. As LD inflates precision spuriously, a purified set of items free of LD was crucial.

Thus, nine CETSCALE items (five of which are part of the ten-item short version) were further scrutinised. Like in Experiment 1, the disagree-to-agree format displayed poor fit when comparing expected scores and actual scores ($\chi^2 = 114.3$, $p = 0.009$). In contrast, fit of the agree-to-disagree format data was satisfactory ($\chi^2 = 96.8$, $p = 0.11$). However, in Experiment 2, there was no difference in the implied unit of measurement as both the standard deviation of the item locations and the standard deviation of person measures were the same for both response scale formats. Likewise, no sign of an additive bias was present as person means did not differ. As expected, the random presentation of agree-to-disagree and disagree-to-agree response scales caused confusion resulting in overall misfit (χ^2 123.4, $p = 0.002$). In this case, the precision of measurement clearly decreased as evidenced by a significant shrinking of the scale by 35 per cent (1-0.50/0.77), which implies that respondents found it much harder to discriminate between the categories. The person mean (−0.48) was also biased relative to the response scales in the other conditions. However, after accounting for the difference in the scale unit, there was no relevant mean difference between any of the three conditions. Consequently, the mean difference did not indicate a true additive bias but rather was a function of the different unit of measurement given the location of the sample (see Table III for key results). In terms of traditional methods to investigate satisficing behaviour, the results confirmed those of Experiment 1. In the agree-to-disagree condition, merely 18 (3.3 per cent) of respondents had (just) one missing value. The disagree-to-agree format resulted in 6 (1.0 per cent) respondents skipping (just) one item response. In terms of straightlining, the agree-to-disagree direction generated 24 (4.3 per cent) straightliners, the disagree-to-agree format 19 (3.3 per cent), the small difference being statistically insignificant ($p = 0.38$) and, if anything, favouring the poorer fitting disagree-to-agree data.

In summary, these results of the CETSCALE-data administered online tend to confirm the superiority of the agree-to-disagree format based on item fit, even though no indication of a relevant difference in the unit was present. The difference was definitely less

Table III.
Key results of the analyses of the CETSCALE in Experiment 2: effect of the response scale direction on fit, the unit of measurement, and precision (online)

Response scale Criterion	Condition 1: Agree-To-Disagree	Condition 2: Disagree-To-Agree	Condition 3: Agree-To-Disagree And Disagree-To-Agree At Random
Overall fit (df)	$\chi^2 = 96.78$ (81), $p = 0.11$	$\chi^2 = 114.34$ (81), $p = 0.009$	$\chi^2 = 123.36$ (81), $p = 0.002$
Mean item fit χ^2 (df)	10.75 (9)	12.70 (9)	13.71 (9)
Standard deviation of item locations	0.77	0.77	0.50
Standard deviation of person locations	1.11	1.14	0.74
Standard deviation of person locations/standard deviation of item locations	1.44	1.49	1.49
Relative unit (standard deviation of item locations divided by standard deviation of item locations in reference group *Condition 1)*	1.00	0.99	0.66
Mean of person locations	−0.73	−0.73	−0.48
Adjusted mean of person locations (mean divided by relative unit)	−0.73	−0.74	−0.74
Regression item location estimates on Condition 1	–	b = 0.99, $r^2 = 0.99$	b = 0.64, $r^2 = 0.98$
Items retained	1, 3, 5*, 6*, 8*, 9, 13*, 15, 16* (Shimp and Sharma, 1987, p.283); * indicates item is part of the ten-item version		

pronounced than in the paper-and-pencil condition. As far as the comparison between the online and the paper-and-pencil-administration is concerned, there is no conclusive sign that one way works better than the other. At least, in the online data the response scale direction appears to matter less. However, there is a difference in the unit between the online and the paper-and-pencil mode. Consequently, caution is advised when comparing measures based on different data collection modes.

Part 2: Analysis of AFC-Scale. The initial analysis of the 8-item AFC scale led to a reduction to six items because of one item misfitting the model and another one due to LD. The results in terms of testing for the general scale direction effect (agree-to-disagree versus disagree-to-agree), when presented to the right of the statement, match those of the CETSCALE in Part 1 perfectly. Data based on the agree-to-disagree format ($\chi^2 = 40.0$, $p = 0.11$) fitted considerably better than data from a disagree-to-agree scale ($\chi^2 = 68.0$, $p < 0.0001$). No relevant difference in the measurement unit occurred, though. Hence, the response scale direction effect seems stable across the two latent constructs.

Testing the spatial proximity hypothesis further, the results from response scales presented to the right of the item and beneath were compared. When presented below the statement, the agree-to-disagree format was no longer superior with fit being actually slightly better for the disagree-to-agree format ($\chi^2 = 40.9$, $p = 0.09$) than for the agree-to-disagree version ($\chi^2 = 51.8$, $p = 0.008$). When the response scale is presented below the item, the previously observed scale direction effect dissolves. On the one hand, this corroborates the proposition of spatial proximity being the explanatory mechanism for the scale direction effect. On the other hand, it raises the question what might be causing the slightly better functioning of the items when using the disagree-to-agree response scale presented below the item.

The conclusion that spatial proximity facilitates the response process is further supported by the fact that the agree-to-disagree format to the right of the item ($\chi^2 = 40.0, p = 0.11$) performed not only better than the disagree-to-agree format at the same position, but also better, if only slightly, than either format presented below the item.

Mean comparisons of person measures between the two response scale directions also favour the agree-to-disagree format presented to the right of the item. While there was no additive bias when the response scale was positioned to the right of the item, the person means differed significantly between the agree-to-disagree (higher, i.e. more agreement) and the disagree-to-agree format (lower, i.e. less agreement) when presented below ($p = 0.003$). Thus, there is a tendency towards the left hand side of the scale, which is only effective when proximity plays no role. In summary, the AFC analysis further strengthens the conclusion that the agree-to-disagree format presented to the right of the item appears to be the best option (see Table IV for key results). As with the CETSCALE response data, satisficing can be ruled out for the AFC data based on extremely low frequencies of missing values (6 times one missing for agree-to-disagree to the right, 4 for disagree-to-agree to the right, 9 for agree-to-disagree beneath and 14 for disagree-to-agree beneath) and straightliners (7 or 1.7 per cent for agree-to-disagree and for disagree-to-agree when to the right, 4 or 1.0 per cent for either version when beneath).

Discussion

Experiment 2 confirms the findings in Experiment 1 insofar as the agree-to-disagree scale also works better than the disagree-to-agree scale in the domain of online-administered surveys. It also shows that the effect replicates across different constructs. Spatial proximity

Response scale Criterion	Condition 1: agree-to-disagree right side of item	Condition 2: disagree-to-agree right side of item	Condition 3: agree-to-disagree beneath item	Condition 4: disagree-to-agree beneath item
Overall fit (df = 30)	$\chi^2 = 39.95$, $p = 0.11$	$\chi^2 = 67.98$, $p < 0.0001$	$\chi^2 = 51.84$, $p = 0.008$	$\chi^2 = 40.86$, $p = 0.09$
Mean of item fit χ^2 (df)	6.66 (5)	11.33 (5)	8.64 (5)	6.81 (5)
Standard deviation of item locations	0.55	0.57	0.61	0.52
Standard deviation of person locations	0.92	0.96	1.03	0.93
Standard deviation of person locations/standard deviation of item locations	1.70	1.70	1.70	1.78
Relative unit (standard deviation of item locations divided by standard deviation of item locations in reference group *Condition 1)*	1.00	1.04	1.11	0.96
Mean of person locations	0.67	0.66	0.82	0.44
Adjusted mean of person locations (mean divided by relative unit)	0.67	0.64	0.74	0.46
Regression item location estimates on Condition 1	–	b = 1.03, $r^2 = 1.00$	b = 1.11, $r^2 = 0.99$	–
Regression item location estimates on Condition 3	–	–	–	b = 0.92, $r^2 = 0.99$

Table IV.
Key results of the analyses of the AFC-scale in Experiment 2: effect of the response scale direction and the positioning of the scale on fit, the unit of measurement and precision (online)

of the statement and the agreement-pole of the scale seems to be the most (psycho-)logical way to present response categories. When no such proximity exists, as it is the case when the response categories are placed below the item, respondents are prone to being biased towards the left hand side. There are undoubtedly other factors which could play a role. The conclusion that agree-to-disagree is superior is tentative only. It should be understood as a default recommendation, if a decision has to be made and no quantitative pre-test of different formats with small samples is feasible due to, e.g. time or financial restrictions. The preferred procedure would be to investigate the functioning of the response scale empirically in a pre-study before large samples are drawn, as the nature of the construct might have an impact. As the effects and their detection are quite subtle, more sophisticated approaches such as the RM, which separates item and person properties and explicitly parameterises the thresholds between response categories, appear to be indispensable.

In Experiment 3, eye-tracking is applied to test whether the response scale direction effect is also reflected in the reading patterns of respondents. In doing so, information is gathered on whether the response scale direction effect is consciously perceived as being disturbing or whether it is a phenomenon that remains at the unconscious level.

Experiment 3

In Experiment 3, eye-tracking as a physiological method is applied to investigate reading patters of respondents when using response scales of different direction in a computer-administered online survey. Based on the psychometric results, different gaze motions were expected. The higher efforts implied by the disagree-to-agree version should increase the time needed to make sense out of the visual stimuli (items and response scales). Consequently, more fixation counts and an enhanced fixation length for this response scale format were expected.

Method

A total of 30 right-handed participants (convenience sample, approximately half students, 11 men, 19 women, mean age 27.7, all having normal or corrected to normal vision) took part in the study. Tobii-eye-tracking technology was used which captures gaze parameters using the corneal reflection technique (Eizenman *et al.*, 1984), which records the visible reflections of a light source on the cornea. Based on geometrical features of these reflections, the gaze direction is calculated (www.tobiipro.com). Respondents were placed in front of a computer screen including the eye-tracker and asked to fill in a survey comprising the CETSCALE (Shimp and Sharma, 1987). Based on random assignment, the participants either got the seven-point-rating scale in the agree-to-disagree or in the disagree-to-agree format. The answer format was placed to the right of the statement. The respective verbal labels of the scale poles were defined as rectangular areas of interest (AOIs). Fixation count and fixation length data, which are associated with cognitive effort and information processing (Just and Carpenter, 1980), were collected. While a longer fixation time suggests a more complex cognitive process, it does not necessarily correspond to subjective impressions of the respondent. Therefore, qualitative post-experimental interviews were administered to gain insights on how the participants subjectively experienced the task of completing the survey and to what extent the response scale direction effect is perceived consciously.

Results

As expected, significantly more fixations were found on the disagree-pole in the group that got the disagree-to-agree format compared to those who got the agree-to-disagree format. This difference could also be substantiated for fixation length, although at a marginally

significant level only. Fixation counts and length comparing the two agree-poles across formats did not yield any significant differences (Table V). Regarding the average time to complete the survey, no significant difference between the two conditions was found. Also regarding reading patterns of the items (statements), no significant differences regarding fixation counts and length were detected.

Furthermore differences between the orientation phase, up to the first response, and the actual response phase were examined. While fixation counts and fixation length of the disagree-pole were almost identical in the two conditions measures during the orientation phase, significantly higher values for both parameters were observed once the first item had been answered. These results indicate that the scale direction effect is not a transitory phenomenon during the phase in which respondents familiarise themselves with the scale. Rather, it occurs throughout the survey introducing unnecessary cognitive extra load for the respondents. The additional cognitive effort might impede respondents allocating appropriate cognitive resources to thinking about their response, resulting in reduced precision and suboptimal data quality.

Discussion

The results of Experiment 3 are compatible with the preceding psychometric analyses suggesting that the disagree-to-agree format adds additional cognitive burden to the participants. Respondents in the disagree-to-agree condition are occupied with handling the response scale instead of fully concentrating on the content of the survey. In principle, respondents in the disagree-to-agree condition could have spent more time in total, thus dedicating the same time to the processing of the item and considering their response as respondents in the agree-to-disagree condition and thereby compensating for the longer time they had to spend on assuring themselves of the response scale. However, they did not, which explains why the quality of the responses in the disagree-to-agree condition was generally poorer.

Qualitative interviews were conducted after the eye-tracking experiment to reveal to what extent respondents consciously struggled with the disagree-to-agree format. Twenty-five out of 30 respondents recalled the response scale format correctly. When asked how they perceived the response scale, 22 said it was okay the way it was, three said the response scale would not matter to them at all, three said agree-to-disagree would have been better than disagree-to-agree and two said the reverse. Thus, respondents overall did not exhibit the slightest preference for either format suggesting that the response scale effects detected throughout our experiments operate at an unconscious level. Further research using bigger samples might provide more robust findings. Physiological data for response scales positioned below the item would be beneficial, too.

Parameter (mean-levels)	Agree-to-disagree		Disagree-to-agree		
	Agree pole	Disagree pole	Disagree pole	Agree pole	
Fixation count	2.140		5.070		$p < 0.05$
Fixation length (in seconds)	0.869		1.808		$p < 0.10$
Fixation count	2.140			1.790	n.s.
Fixation length (in seconds)	0.869			0.712	n.s.
Fixation count		2.640	5.070		$p < 0.05$
Fixation length (in seconds)		1.023	1.808		$p < 0.10$

Table V. Key results of Experiment 3: effects of the response scale direction on gaze parameters (online)

Conclusions, implications and further research

General conclusions

Despite a long history of research into the direction of the response scale, no coherent picture has emerged with scattered and mixed evidence. At least in part the situation seems to be due to limitations in standard psychometric analyses. The RM for measurement lends itself much better to the psychometric investigation of quite subtle and intricate response scale effects for at least three reasons. First, it provides better possibilities to assess item fit. Second, it parameterises the response scale by specifying threshold parameters. Third, it statistically separates item and person characteristics.

The identified psychometric differences imply that the response scale direction may matter in terms of item fit and precision. Misfit caused by a suboptimal response format may wrongly be attributed to the items. Poorer measurement precision may also reduce effect sizes in substantive studies. It remains to be seen whether one response scale direction proves universally preferable. Meanwhile, experimental pre-studies are strongly advised to inform the response scale design.

Extreme caution is due when data based on different response scale directions are to be compared or merged. As Experiment 2 demonstrates, a difference in the unit of measurement may result in a spurious mean difference. As the presence and the direction of this effect depend only on the relative locations of the samples, a difference in the unit may trigger opposite effects under different circumstances. This may explain the observation in the literature that sometimes a particular scale direction implies a positive bias, sometimes a negative, and sometimes none at all.

In terms of the responsible mechanism, the experiments support the spatial proximity heuristic. When presenting the response scale to the right of the item, there is no indication of an additive bias and, thus, no need to invoke primacy effects or anchoring heuristics. With spatial proximity no longer present when positioning the response scale below the item, the differences between the two response scale directions become small. However, the additive bias occurring in Experiment 2 suggests presenting the response scale to the right of the item seems recommendable.

Another aspect worth considering is response burden. Social researchers ought to minimise the impact their research activities might have on study participants and avoid any unnecessary burden. The eye-tracking study provides clear indication that a suboptimal response format might confuse participants and increase response burden. It is safe to assume that their struggle with the response scale is a direct cause of the poorer data quality observed. Thus, striving for the best possible response scale format is likely to result in a win-win situation from which both the participants and the researchers can benefit.

As a general conclusion, scholars are strongly urged to pay more attention to the response scale and its characteristics. The present study illustrates how the RM can be used to identify effects of various manipulations of the response scale, such as a different number of the response categories, their verbalisation, or the use of numerical labels. The response categories should undergo the same rigorous procedure that is applied to the item generation including a literature review, expert advice, qualitative interviews (cognitive debriefing) with consumers, and ultimately, psychometric analysis.

The use of non-representative samples may be a limitation of the present study. However, given the study objectives, this restriction is not deemed critical as all experiments were conducted using random assignment and comparable samples. The emphasis lies on the proof of concept with respect to the suitability of the methodology and the proposed mechanism.

With respect to Experiment 2, one might object that presenting the response scale below the item still involves proximity insofar as after reading the item the respondent starts reading the response scale at its left end. However, first of all, this type of proximity is temporal in nature, while the proposed mechanism involves quite literally spatial proximity. Second, it is suggested that spatial proximity does not refer to comprehending the item and the associated response scale. Rather, it is supposed to matter when it comes to providing the response. At that stage all response options are equidistant to the item when placed below the item.

Academic and managerial implications

When using published measurement instruments, any available evidence presented in support of the response scale functioning, as sparse as it may be, should be carefully considered. When carrying out recurring measurements, such as in customer satisfaction monitoring or in longitudinal studies, it is pivotal to use the same response format throughout the study. Changes to the format require complicated psychometric analyses and may result in incomparable metrics. Cross-sectional comparisons based on data arising from different response formats, for example in benchmarking across industries, should be carried out with caution.

Further research

Ideally, every study applying a measurement instrument should estimate and report Rasch item and threshold locations along with a detailed account of the design of the instrument and the conditions of data collection. This would allow for conclusions in terms of the unit of measurement in psychometric meta-analyses. It remains to be seen whether the superiority of the agree-to-disagree format generalises to different instruments, reversed items (Swain *et al.*, 2008), types of response scales (number of categories, verbal or numerical labels), populations, for example in terms of culture (Lee *et al.*, 2002; Craig and Douglas, 2005), and the type of script (left-to-right versus right-to-left).

Remarkably, the detrimental effect of a suboptimal response scale seems to operate at an unconscious level. Further research may investigate whether the response scale enhances the level of arousal at an implicit level using physiological methods, such as skin conductance (Walla *et al.*, 2011). The startle reflex modulation (Koller and Walla, 2012) lends itself as a method to examine objectively underlying affective processes.

Finally, while the methodology presented in this study allows for a thorough investigation of the unit of measurement, it must not be overlooked that the unit is still implicit. It is up to conceptual work to lay ground for developing explicit, tangible units of measurement in the social sciences.

References

Andrich, D. (1978), "Application of a psychometric rating model to ordered categories which are scored with successive integers", *Applied Psychological Measurement*, Vol. 2 No. 4, pp. 581-594.

Andrich, D. (1988), "A general form of rasch's extended logistic model for partial credit scoring", *Applied Measurement in Education*, Vol. 1 No. 4, pp. 363-378.

Andrich, D. (2002), "Implications and applications of modern test theory in the context of outcomes based education", *Studies in Educational Evaluation*, Vol. 28 No. 2, pp. 103-121.

Andrich, D. (2011), "Rating scales and rasch measurement", *Expert Review in Pharmacoeconomics and Outcomes Research*, Vol. 1 No. 5, pp. 571-585.

Andrich, D., Sheridan, B.S. and Luo, G. (2009-2012), "Rumm2030: rasch unidimensional measurement models", *Computer Software*, RUMM Laboratory Perth, Western Australia.

Baumgartner, H. and Steenkamp, J.-B.E.M. (2001), "Response styles in marketing research: a cross-national investigation", *Journal of Marketing Research*, Vol. 38 No. 2, pp. 143-156.

BIPM, I., IFCC, I., ISO, I. IUPAP. and OIML, (2008), "Evaluation of measurement data–guide to the expression of uncertainty in measurement", Joint Committee for Guides in Metrology, Technical Report No. JCGM 100.

Bordé, C.J. (2005), "Base units of the SI, fundamental constants and modern quantum physics", *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, Vol. 363 No. 1834, pp. 2177-2201.

Bradburn, N., Sudman, S. and Wansink, B. (2004), *Asking Questions – the Definitive Guide to Questionnaire Design – for Market Research, Political Polls, and Social and Health Questionnaires*, Jossey-Bass, San Francisco, CA

Clarke, I.I.I. (2001), "Extreme response style in cross-cultural research", *International Marketing Review*, Vol. 18 No. 3, pp. 301-324.

Cox, E.P. III, (1980), "The optimal number of response alternatives for a scale: a review", *Journal of Marketing Research*, Vol. 17 No. 4, pp. 407-422.

Craig, C.S. and Douglas, S.P. (2005), *International Marketing Research*, John Wiley and Sons, Hoboken, NJ.

Deutskens, E., de Ruyter, K. and Wetzels, M. (2006), "An assessment of equivalence between online and mail surveys in service research", *Journal of Service Research*, Vol. 8 No. 4, pp. 346-355.

Dickson, J. and Albaum, G. (1975), "Effects of polarity of semantic differential scales in consumer research", *Advances for Consumer Research*, Vol. 2, pp. 507-514.

Dolnicar, S., Laesser, C. and Matus, K. (2009), "Online versus paper – format effects in tourism surveys", *Journal of Travel Research*, Vol. 47 No. 3, pp. 295-316.

Eizenman, M., Frecker, R. and Hallett, P. (1984), "Precise non-contacting measurement of eye movements using the corneal reflex", *Vision Research*, Vol. 24 No. 2, pp. 167-174.

Embretson, S.E. and Reise, S.P. (2013), *Item Response Theory*, Psychology Press, London.

Ewing, M.T., Salzberger, T. and Sinkovics, R.R. (2005), "An alternate approach to assessing cross-cultural measurement equivalence in advertising research", *Journal of Advertising*, Vol. 34 No. 1, pp. 17-36.

Fanning, E. (2005), "Formatting a paper-based survey questionnaire: best practices", *Practical Assessment Research and Evaluation*, Vol. 10 No. 12, pp. 1-14.

Fischer, G.H. (1995), "Derivations of the rasch model", in Fischer, G.H. and Molenaar, I.W. (Eds), *Rasch Models, Foundations Recent Developments, and Applications*, Springer, Berlin, Germany, pp. 15-38.

Friedman, H.H., Friedman, L.W. and Gluck, B. (1988), "The effects of scale-checking styles on responses to a semantic differential scale", *Journal of the Marketing Research Society*, Vol. 30 No. 4, pp. 477-481.

Friedman, H.H., Herskovitz, P.J. and 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.

Galesic, M. (2006), "Dropouts on the web: Effects of interest and burden experienced during an online survey", *Journal of Official Statistics*, Vol. 22 No. 2, p. 313.

Ganglmair-Wooliscroft, A. and Wooliscroft, B. (2013), "A cross-cultural application of the affective response to consumption scale: investigating US-American and Austrian passengers on long-haul flights", *Journal of Business Research*, Vol. 66 No. 6, pp. 765-770.

Ganglmair-Wooliscroft, A. and Wooliscroft, B. (2016), "Diffusion of innovation: the case of ethical tourism behavior", *Journal of Business Research*, Vol. 69 No. 8, pp. 2711-2720.

Greenleaf, E.A. (1992), "Improving rating scale measures by detecting and correcting bias components in some response styles", *Journal of Marketing Research*, Vol. 29 No. 2, pp. 176-188.

Humphry, S.M. (2005), "*Maintaining a common arbitrary unit in social measurement"*, Dissertation, School of Education, Murdoch University, Perth, WA.

Humphry, S.M. and Andrich, D. (2008), "Understanding the unit in the rasch model", *Journal of Applied Measurement*, Vol. 9 No. 3, pp. 249-264.

Jiménez-Guerrero, J.F., Gázquez-Abad, J.C. and del Carmen Linares-Agüera, E. (2014), "Using standard CETSCALE and other adapted versions of the scale for measuring consumers' ethnocentric tendencies: an analysis of dimensionality", *BRQ Business Research Quarterly*, Vol. 17 No. 3, pp. 174-190.

Just, M.A. and Carpenter, P.A. (1980), "A theory of reading: from eye fixations to comprehension", *Psychological Review*, Vol. 87 No. 4, pp. 329-354.

Koller, M. and Walla, P. (2012), "Measuring affective information processing in information systems and consumer research - Introducing startle reflex modulation", *ICIS Proceedings, Association for Information Systems (AIS)*, pp. 1-16.

Krosnick, J.A. (1999), "Survey research", *Annual Review of Psychology*, Vol. 50 No. 1, pp. 537-567.

Lee, J.W., Jones, P.S., Mineyama, Y. and Zhang, X.E. (2002), "Cultural differences in responses to a likert scale", *Research in Nursing and Health*, Vol. 25 No. 4, pp. 295-306.

Liu, M. and Keusch, F. (2017), "Effects of scale direction on response style of ordinal rating scale", *Journal of Official Statistics*, Vol. 33 No. 1, pp. 137-154.

Lord, F.M. and Novick, M.R. (Eds) (1968), *Statistical Theories of Mental Test Scores*, Addison-Wesley, Boston, MA.

Lozano, L.M., García-Cueto, E. and Muñiz, J. (2008), "Effect of the number of response categories on the reliability and validity of rating scales", *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, Vol. 4 No. 2, pp. 73-79.

MacKenzie, S.B., Podsakoff, P.M. and Podsakoff, N.P. (2011), "Construct measurement and validation procedures in MIS and behavioral research: Integrating new and existing techniques", *MIS Quarterly*, Vol. 3 No. 2, pp. 293-334.

Marais, I. (2013), "Local dependence", in Christensen, K.B., Kreiner, S. and Mesbah, M. (Eds) *Rasch Models in Health*, Wiley-ISTE Ltd, London, UK and Hoboken, NJ, pp. 111-130.

Marais, I. and Andrich, D. (2008), "Effects of varying magnitude and patterns of local dependence in the unidimensional rasch model", *Journal of Applied Measurement*, Vol. 9 No. 2, pp. 1-20.

Mathews, C.O. (1929), "The effect of printed response words on an interest questionnaire", *Journal of Educational Psychology*, Vol. 20 No. 2, pp. 128-134.

Mead, A.D. and Drasgow, F. (1993), "Equivalence of computerized and paper-and-pencil cognitive ability tests: a Meta-analysis", *Psychological Bulletin*, Vol. 11 No. 3, pp. 449-458.

Menold, N. and Tausch, A. (2016), "Measurement of latent variables with different rating scales: testing reliability and measurement equivalence by varying the verbalization and number of categories", *Sociological Methods and Research*, Vol. 45 No. 4, pp. 678-699.

Parasuraman, A., Zeithaml, V.A. and Berry, L.L. (1998), "Alternative scales for measuring service quality: a comparative assessment based on psychometric and diagnostic criteria", in Bruhn, M. and Meffert, H. (Eds), *Handbuch Dienstleistungsmanagement*, Gabler Verlag, Wiesbaden, Germany, pp. 449-482.

Pendrill, L., Cano, S., Barbic, S. and Fisher, W.P. Jr, (2017), "Patient-centred outcome metrology for healthcare decision-making", in *Joint IMEKO TC1-TC7-TC13 Symposium:"Measurement Science Challenges in Natural and Social Sciences*", Rio de Janeiro, Brazil, July 31 to August 3, 2017.

Pendrill, L.R. and Fisher, W.P. Jr, (2013), "Quantifying human response: linking metrological and psychometric characterisations of man as a measurement instrument", *Journal of Physics: Conference Series*, Vol. 459 No. 1, p. 012057.

Rammstedt, B. and Krebs, D. (2007), "Does response scale format affect the answering of personality scales? Assessing the big five dimensions of personality with different response scales in", *a Dependent Sample", European Journal of Psychological Assessment*, Vol. 23 No. 1, pp. 32-38.

Ramón Oreja-Rodríguez, J. and Yanes-Estévez, V. (2010), "Environmental scanning: dynamism with rack and stack from rasch model", *Management Decision*, Vol. 48 No. 2, pp. 260-276.

Rasch, G. (1960), *Probabilistic Models for Some Intelligence and Attainment Tests*, Danish Institute for Educational Research, Copenhagen, expanded edition (1980).

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

Raykov, T. and Calantone, R.J. (2014), "The utility of item response modeling in marketing research", *Journal of the Academy of Marketing Science*, Vol. 42 No. 4, pp. 337-360.

Rea, L.M. and Parker, R.A. (2014), *Designing and Conducting Survey Research: A Comprehensive Guide*, John Wiley and Sons, Hoboken, NJ.

Salzberger, T. (2007), "The impact of global warming on consumer behaviour – first findings from an empirical study", in *Proceedings of the Australian and New Zealand Marketing Academy Conference*, School of Business, University of Otago, Dunedin.

Salzberger, T. (2009), *Measurement in Marketing Research: An Alternative Framework*, Edward Elgar, Cheltenham Glos and Northampton.

Salzberger, T. and Koller, M. (2013), "Towards a new paradigm of measurement in marketing", *Journal of Business Research*, Vol. 66 No. 9, pp. 1307-1317.

Salzberger, T., Newton, F.J. and Ewing, M.T. (2014), "Detecting gender item bias and differential manifest response behavior: a rasch-based solution", *Journal of Business Research*, Vol. 67 No. 4, pp. 598-607.

Salzberger, T., Sinkovics, R. and Holzmüller, H. (1997), "Problems of equivalence in cross-cultural marketing research", in *Proceedings of the 1997 Academy of Marketing Science (AMS) Annual Conference*, Springer, Berlin, Germany, pp. 74-78.

Sheluga, D., Jacoby, J. and Major, B. (1978), "Whether to agree-disagree or disagree-agree: the effects of anchor order on item response", *Advances of Consumer Research*, Vol. 5 No. 1, pp. 109-113.

Shimp, T. and Sharma, S. (1987), "Consumer ethnocentrism: construction and validation of the CETSCALE", *Journal of Marketing Research*, Vol. 24 No. 3, pp. 280-289.

Sinkovics, R.R. (1999), "Ethnozentrismus und konsumentenverhalten", [*Ethnocentrism and Consumer Behaviour]*, Deutscher Universitätsverlag, Wiesbaden.

Steenkamp, J.B.E., De Jong, M.G. and Baumgartner, H. (2010), "Socially desirable response tendencies in survey research", *Journal of Marketing Research*, Vol. 47 No. 2, pp. 199-214.

Swain, S.D., Weathers, D. and Niedrich, R.W. (2008), "Assessing three sources of misresponse to reversed likert items", *Journal of Marketing Research*, Vol. 45 No. 1, pp. 116-131.

Sweeney, J.C., Danaher, T.S. and McColl-Kennedy, J.R. (2015), "Customer effort in value cocreation activities: improving quality", *Of Life and Behavioral Intentions of Health Care Customers", Journal of Service Research*, Vol. 18 No. 3, pp. 318-335.

Tourangeau, R., Couper, M.P. and Conrad, F. (2004), "Spacing, position, and order, interpretive heuristics for visual features of survey questions", *Public Opinion Quarterly*, Vol. 68 No. 3, pp. 368-393.

Traub, R.E. (1994), "*Reliability for the Social Sciences: Theory and Applications*", Vol. 3, SAGE Publications, Thousand Oaks.

Tversky, A. and Kahneman, D. (1974), "Judgment under uncertainty: heuristics and biases", *Science*, Vol. 185 No. 4157, pp. 1124-1131.

Voss, K.E., Stem, D.E., Jr and Fotopoulos, S. (2000), "A comment on the relationship between coefficient alpha and scale characteristics", *Marketing Letters*, Vol. 11 No. 2, pp. 177-191.

Walla, P., Brenner, G. and Koller, M. (2011), "Objective measures of emotion related to Brand attitude: a new way to quantify Emotion-Related aspects relevant to marketing", *PLoS ONE*, Vol. 6 No. 11, pp. 1-7.

Weijters, B., Cabooter, E. and Schillewaert, N. (2010), "The effect of rating scale format on response styles: the number of response categories and response category labels", *International Journal of Research in Marketing*, Vol. 27 No. 3, pp. 236-247.

Weng, L. (2004), "Impact of the number of response categories and anchor labels on coefficient alpha and test-retest reliability", *Educational and Psychological Measurement*, Vol. 64 No. 6, pp. 956-972.

Wong, C.S., Tam, K.C., Fung, M.Y. and Wan, K. (1993), "Differences between odd and even number of response scale: some empirical evidence", *Chinese Journal of Psychology*, Vol. 35 No. 2, pp. 75-86.

Wright, B.D. and Mok, M. (2000), "Understanding rasch measurement: rasch models overview", *Journal of Applied Measurement*, Vol. 1 No. 1, pp. 83-106.

Yan, T. and Keusch, F. (2015), "The effects of the direction of rating scales on survey responses in a telephone survey", *Public Opinion Quarterly*, Vol. 79 No. 1, pp. 145-165.

Yen, W.M. (1984), "Effects of local item dependence on the fit and equating performance of the three-parameter logistic model", *Applied Psychological Measurement*, Vol. 8 No. 2, pp. 125-145.

Further reading

www.tobiipro.com [Website], " How do tobii eye trackers work?", accessed 27 February 2018, CET 14:12, Tobii AB, Danderyd, Stockholm.

Appendix

Items of affective concern (AFC, Salzberger, 2007), working translations only

- Climate change is a serious problem for humanity.
- I am annoyed with people who use their car even for short distances.
- It bothers me when people waste energy senselessly.
- If people do not properly separate their trash, it makes me angry.
- When I look at how the glaciers melt, it hurts.
- If I once again bought a product that harms the environment, it burdens me afterwards.
- I find it annoying when the air is so polluted.
- Ignorance of environmental problems offends me.
- I am sad that we may leave our descendants with lasting environmental damage.

Corresponding author

Thomas Salzberger can be contacted at: thomas.salzberger@wu.ac.at