gesis Leibniz-Institut für Sozialwissenschaften

GESIS Survey Guidelines

Design of Rating Scales in Questionnaires

Natalja Menold & Kathrin Bogner

December 2016, Version 2.0

Abstract

Rating scales are among the most important and most frequently used instruments in social science data collection. There is an extensive body of methodological research on the design and (psycho)metric properties of rating scales. In this contribution we address the individual scale-related aspects of questionnaire construction. In each case we provide a brief overview of the current state of research and practical experience, and – where possible – offer design recommendations.

Citation

Menold, N., & Bogner, K. (2016). Design of Rating Scales in Questionnaires. *GESIS Survey Guidelines*. Mannheim, Germany: GESIS – Leibniz Institute for the Social Sciences. doi: 10.15465/gesis-sg_en_015

1. Introduction

Since their introduction by Thurstone (1929) and Likert (1932) in the early days of social science research in the late 1920s and early 1930s, rating scales have been among the most important and most frequently used instruments in social science data collection. A rating scale is a continuum (e.g., agreement, intensity, frequency, satisfaction) with the help of which different characteristics and phenomena can be measured in questionnaires. Respondents evaluate the content of questions and items by marking the appropriate category of the rating scale. For example, the European Social Survey (ESS) question "All things considered, how satisfied are you with your life as a whole nowadays?" has an 11-point rating scale ranging from 0 (*extremely dissatisfied*) to 10 (*extremely satisfied*). Rating scales are the subject of the present contribution. They are just one way of documenting responses in questionnaires. Other types of predefined response options, for example lists of nominal categories, are also used in questionnaires. However, they are not addressed here (see also the *GESIS Survey Guidelines* contribution "Question Wording," Lenzner & Menold, 2016).

Respondents' answers are perceived, first, as a function of two fundamental characteristics of rating scales (Parducci, 1983): (1) range, which is delimited by the poles of the scale, and (2) frequency, which is determined by the number of response categories. The number of categories and the labels of the scale endpoints are thus of fundamental importance for understanding the continuum to be measured. As one context of question response, rating scales can lead to both desirable effects, such as enhancing the intended understanding of the range and frequency, and undesirable effects, such as acquiescence (i.e., the tendency to agree with items regardless of their content). These undesirable effects can be reduced by designing rating scales in a certain way. Not only the number and labelling of response categories but also graphical features of rating scales, such as scale orientation or the use of colours, type fonts, and shading, can influence the way in which rating scales are understood (*visual design*, e.g., Tourangeau, Couper & Conrad, 2004; 2007). The objective of rating scale design is to motivate respondents to answer the questions in a diligent way (so-called *optimising;* Krosnick & Alwin, 1987). Generally, the task of answering questions should not be too complex or difficult, nor should it unnecessarily tempt the respondents to reduce their cognitive burden (so-called *satisficing*; Krosnick & Alwin, 1987).

Extensive methodological research on the design of rating scales and their (psycho)metric properties has been conducted – mainly between the 1960s and the 1980s. More recent studies have been carried out mostly in the context of the visual design approach. Also worthy of mention are recent multitrait-multimethod (MTMM) investigations that used structural equation models (e.g., Saris & Gallhofer, 2007). Design aspects of rating scales, such as (a) the number of categories, (b) the inclusion of a scale midpoint (c) the use of verbal or numerical labels, (d) scale orientation, and (e) scale polarity, have been investigated (see systematic reviews by Krosnick & Fabrigar, 1997 or Menold & Bogner, 2012). Other research has compared the use of item-specific scales – for example, importance, frequency, and satisfaction – with universally applicable agree/disagree scales (Likert-type scales). (This is discussed in more detail in Section 8 below).

The effects of rating scales have been investigated mainly in relation to psychometric quality criteria. These criteria include reliability (the precision of a measurement; see also the *GESIS Survey Guidelines* contribution "Reliability", Danner, 2016) and validity (an indication of the extent to which statements about the concepts to be measured can be made on the basis of the measurement results). Moreover, systematic measurement error in the form of extreme response style, middle response style, and item nonresponse, respondent preferences, and the difficulties they experience when answering survey questions have been studied.

In what follows, we address the individual rating-scale-related aspects of questionnaire design. In each case, we briefly outline the current state of research and – to the extent possible – offer design recommendations.

2. Number of response categories

The number of response categories is a very important scale characteristic because – following Parducci's (1983) range-frequency model (see Section 1 above) – it determines the degree of differentiation of the rating scale and thus the ease of understanding of the continuum in question.

In systematic reviews, Krosnick and colleagues (Krosnick & Fabrigar, 1997; Krosnick & Presser, 2010) came to the conclusion that an optimal measurement – in terms of reliability, validity, and degree of differentiation – could be achieved with five to seven categories. Respondents also preferred scales of this length (Krosnick & Fabrigar, 1997). This finding is explained by the fact that using too many categories reduces the clarity of meaning of the individual categories, which makes it more difficult for respondents to answer the question. On the other hand, if a small number of categories are used, the rating scale is not sufficiently differentiated. Menold and colleagues (Menold & Tausch, 2015; Menold & Kemper, 2015) showed that the reliability of scales with five categories may also differ from that of scales with seven categories. However, these results were dependent on the other characteristics of rating scales, particularly, their verbal and/or numerical labelling.

Other studies have found a linear relationship between the number of categories and psychometric quality criteria. In other words, as the number of categories increased, so, too, did the measurement quality (e.g., Saris & Gallhofer, 2007; Pajares, Hartley, & Vahante, 2001; Preston & Colman, 2000). The maximum number of categories tested in these studies varied between 10, 11, and 100. In practice, however, the established rule of thumb with regard to scale length appears to be five to seven points, especially because such scales are easier to verbally label (see next section).

Conclusion and recommendation: In line with the majority of research studies, which support the "five to seven points" rule, we recommend this number of categories. In addition, one should also take into account the labelling and verbalisation of rating scales. In certain cases, individual characteristics of questions or the use of special data analysis methods may lead researchers to decide to use more than seven categories.

3. Category labels

When it comes to using category labels, one must decide whether to verbally label only the endpoint categories, which delimit the range of the rating scale, or whether verbal labels should be used for each category. In addition to verbal labels, numerical labels are very frequently used in rating scales.

Systematic reviews have concluded that verbally labelling all the rating scale categories increases test-retest reliability and validity (Maitland, 2009). Saris and Gallhofer (2007) and Menold, Kaczmirek, Lenzner, and Neusar (2014) found that fully verbalised response scales also increased cross-sectional reliability (split-half method in the study by Menold et al., 2014). Menold and colleagues (Menold & Tausch, 2015; Menold & Kemper, 2015) found that measurement quality was also dependent on the number of categories. When five-category, fully verbalised rating scales were used, measurement problems were sometimes apparent, while they were reduced in the case of seven-category, fully verbalised rating scales.

Moreover, various studies have shown that respondents prefer fully verbalised rating scales (e.g., Wallsten, Budescu, & Zwick, 1993; Zaller, 1988). Furthermore, verbally labelling all the points in a rating scale has been shown to reduce undesirable effects of other visual elements in questionnaires (e.g., Toepoel & Dillman, 2011). The positive effect of fully verbalising rating scales is explained by the fact that it makes the meanings of the categories clearer – compared to scales that do not use verbal labels for all categories. People with low and moderate formal education especially benefit from full verbalisation (Krosnick & Fabrigar, 1997).

Numerical labels, which are very frequently used in surveys, are – in theory – less favourable than verbal labels. On the one hand they may mean different things to different people (e.g., lucky or unlucky numbers, academic grades), and these meanings may be incongruent with those of the categories. On the other hand, it is neither very natural nor self-evident to describe oneself or others numerically (Krosnick & Fabrigar, 1997). Studies reported by Krosnick and Fabrigar and other studies (Windschitl & Wells, 1996; Christian, Parsons, & Dillman, 2009; Menold & Kemper, 2015) support this assumption. Furthermore, as the results by Menold and Kemper (2015) show, one should be particularly careful when combining numerical and verbal labels in one rating scale.

Verbalised rating scales should meet the following requirements: First, the verbal labels should be precise. Second, the rating scales should be balanced. Balanced rating scales are symmetrical, that is, they have the same number of positive and negative categories. Third, the verbal labels should be generally comprehensible, or universal. And fourth, the rating scale categories should suggest apparently equidistant ranges between the categories. Designing a rating scale in this way is by no means a trivial task. For example, Pollack, Friedman, and Presby (1990) showed in relation to universal comprehensibility that the emotional colouring and extremeness of verbal labels in rating scales had an effect on response behaviour. Moreover, certain verbal labels have been shown to contribute more to negative skew than others (French-Lazovik & Gibson, 1984). And Worcester and Burns (1975) showed that antonyms were not necessarily perceived as such in the order in which they appeared on rating scales. Furthermore, the percentage values that respondents associated with verbal quantifiers (e.g., “rare,” “unlikely,” “possible”) varied very strongly across different studies (Theil, 2002), which points to the fact that these quantifiers are understood differently.

In the 1980s and 1990s, a number of studies were conducted to develop universally applicable verbal labels for a number of different continua – for example, frequency, evaluation, and probability – for different word types, such as adjectives and adverbs. Most of these studies were conducted in the English-speaking area (for a review, see Clark, 1990), However, Rohrmann (1978) proposed German-language verbal labels for various continua.

Conclusion and recommendation:

Current research findings support the use of fully verbalised rating scales more than rating scales (e.g., with numerical labels) in which only the endpoints are verbally labelled Fully verbalised scales can especially benefit people with low and moderate formal education. The above-mentioned studies on the design of fully verbalised rating scales can be referred to when selecting verbal labels. The use of full verbal labelling combined with a moderate number of response categories, in particular seven (see above), would appear to be a practicable approach.

4. Scale polarity

The term *scale polarity* refers to the distinction between unipolar and bipolar rating scales. Bipolar rating scales comprise two opposite continua, such as positive/negative and agree/disagree; unipolar

rating scales consist of a continuum from a low to a high level, for example not satisfied at all/very satisfied.

Saris and Gallhofer (2007) investigated the fit between the polarity of dimensions and that of rating scales. They defined a dimension as unipolar if an opposite dimension was not conceivable. For example, linguistic opposites exist for the level of satisfaction or happiness (dissatisfied/satisfied, unhappy/happy). However, no such opposites exist for the *frequency* dimension – that is, one can distinguish only between *never* and *always*. Accordingly, Saris and Gallhofer (2007) concluded that unipolar rating scales should be used for dimensions such as frequency, and bipolar rating scales should be used for the dimensions of satisfaction and happiness. However, they found that such a fit did not have any effect on the quality criteria.

There is, however, no uniform definition of scale polarity in the literature. For example, Krosnick and Fabrigar (1997) considered *importance* to be a unipolar dimension whereas, following the definition used by Saris and Gallhofer (2007), it would be classified as a bipolar dimension. Other authors have defined polarity as the use of numerical labels: negative and positive numerical values are found in bipolar rating scales; numerical values running from zero upwards are found in unipolar rating scales (Moors, Kieruj & Vermunt, 2014). In relation to the use of negative numerical values, various studies have shown that respondents avoid the negative side of the scale and produce more positive answers (Schwarz et al., 1991; Schaeffer & Barker, 1995).

The fact that many dimensions can be realised both as unipolar and bipolar scales raises the question of the fundamental advantages of bipolar rating scales. Krebs (2012) and Menold and Raykov (2015) showed that the use of unipolar or bipolar labels in rating scales may influence the psychometric properties of multi-item measurements and the measured values of the latent variables. However, the findings of Krebs (2012) and Menold and Raykov (2015) are contradictory, so that more research is needed before a final recommendation can be given as to whether unipolar or bipolar rating scales should be used.

Conclusion and recommendation: In general, little research has been conducted on the effects of scale polarity. Hence, it is not possible to make unequivocal recommendations in this regard, except to say that negative numerical labels may produce systematic effects – that is, more positive responses – and should therefore be avoided.

5. Scale orientation

Scale orientation refers to the decision whether the lowest/most negative value should be placed at the beginning of the scale and the highest/most positive value should be placed at the end (ascending order), or vice versa (descending order) – for example, “strongly disagree,” “disagree,” “neither agree nor disagree,” “agree,” “strongly agree,” or vice versa.

While stronger response order effects may occur in the case of vertically presented response categories, where the first- or last-presented categories are preferentially chosen (primacy and recency effects; Krosnick & Alwin, 1987; Toepoel, 2008), such effects are only slight when categories are presented horizontally (Tourangeau, Rips, & Rasinski, 2000). In the latter case, the category at the left-hand end of the rating scale is chosen more frequently than the category at the right-hand end, irrespective of scale orientation. This is known as a “general primacy effect”. However, more recent studies found a primacy effect in the case of descending presentation order (Hofmans et al., 2007; Krebs & Hoffmeyer-Zlotnick, 2010). The few studies that have investigated the effect of scale orientation on measurement quality could find no effects (Saris & Gallhofer, 2007; Krebs & Hoffmeyer-Zlotnik, 2010).

Conclusion and recommendation: At present, no strong recommendations regarding the choice of scale orientation for horizontally presented rating scales can be derived from the findings in the literature. As scale orientation has been found to have no effect on psychometric measurement quality criteria, researchers can arrange rating scales either in ascending or descending order.

6. Scale midpoint

When designing response scales, one must decide whether or not to include a scale midpoint. First, the polarity of the rating scale discussed above must be taken into account because, depending on whether a scale is unipolar or bipolar, the middle category expresses a different position on the part of the respondent: In bipolar rating scales, the middle category can express either indifference (neither/nor) or ambivalence (partly/partly; Kaplan, 1972; Dubois & Burns, 1975). This ambiguity renders the interpretation of the middle category in bipolar rating scales more difficult for the respondents and for the researcher. In a unipolar rating scale, the middle category stands for a middle position, which finds expression in labels such as “somewhat true” or “agree to some extent”.

Besides polarity, three further potential sources of error that arise from the inclusion or non-inclusion of a middle category must be weighed against each other:

First, offering a middle alternative may constitute an invitation to those respondents who tend to satisfice. Satisficing respondents are usually people who are poorly motivated or fatigued. They choose the middle category in order to reduce the cognitive burden of answering the questions and not because it corresponds to their actual opinion. However, most respondents tend towards one scale direction, and if a scale midpoint was not included, they would report that opinion (Krosnick, 1991). Various experimental studies have come to the conclusion that including a middle or neutral category increases response non-differentiation and may lead to less thorough responding (Kalton, Robert, & Holt, 1980; Krosnick & Fabrigar, 1997; Schumann & Presser, 1981; Saris & Gallhofer, 2007). Bishop, Oldendick, Tuchfarber, and Bennett (1980) and O’Muircheartaigh, Krosnick, and Helic (1999) found that respondents chose the middle category more often when (a) they did not consider the issue in question to be very important, (b) they were not interested in it, or (c) they did not have a strong opinion on it. However, no relation was found between the frequency of midpoint selection and the respondent’s knowledgeability about the topic (O'Muircheartaigh et al., 1999). Various studies have also confirmed that middle alternative selection was unrelated to educational attainment (Kalton, Roberts & Holt, 1980; Schuman & Presser, 1981; O’Muircheartaigh et al., 1999; Krosnick, Narayan, & Smith, 1996).

Second, besides respondents whose choice of the middle category is the result of satisficing, there are those who do, in fact, have a neutral or moderate attitude towards the issue in question. If these respondents are given a scale that does not have a middle alternative, they are unable to correctly express their neutral or moderate opinion. This therefore raises the question whether these respondents use another response category randomly or systematically, and whether systematic errors occur as a result. O’Muircheartaigh et al. (1999) showed that the inclusion of a middle alternative increased the reliability and validity of scales. Moreover, various studies have come to the conclusion that, when a midpoint is omitted, respondents do not randomly choose another category but rather systematically select a category near the actual midpoint of the scale (Krosnick, 2002; Schumann & Presser, 1981). Therefore, Krosnick and Presser (2010) recommended that a middle category should be offered.

The third potential source of error is the possibility that, for reasons of social desirability, respondents who do not have any opinion on the issue will choose a middle category rather than reporting that they have no opinion. As a result, the proportion of the population who have an opinion on the issue is

overestimated on the basis of the survey data. Moreover, the assumption of ordinality is violated because the middle category no longer represents only neutral or moderate opinions but also the lack of an opinion (e.g., Sturgis, Roberts, & Smith, 2014). In a test-retest study, Kulas, Stachowski, and Haynes (2008) found that respondents often used the middle category as a substitute for a missing "don't know" category. However, this behaviour did not have an effect on the validity and reliability of the personality scales investigated. Therefore, the authors recommended that a middle category should be offered in rating scales. Sturgis et al. (2014) also showed with the help of follow-up probes that a large proportion of those respondents who had chosen the middle category actually had no opinion on the issue in question. The authors termed these middle-category responses "face-saving don't knows". However, in contrast to Kulas et al. (2008), they found that reallocating "face-saving don't knows" to the "don't know" response category significantly altered the distributions of the investigated items. Moreover, their results show that the tendency to select the scale midpoint as a face-saving way of saying "don't know" was more pronounced among respondents who were of the opinion that they should hold, and report, an opinion on important issues. This introduced a systematic error into the data. However, the follow-up probes also revealed that the other group of respondents who chose the middle category did so because they did, in fact, have a neutral attitude to the issue in question. Therefore, Sturgis et al. (2014) recommended that the middle category should be offered in rating scales in order to prevent people with a neutral opinion from being forced to choose a substantively incorrect response.

Conclusion and recommendation: Research findings on the effects of the middle category show that respondents choose it not only – as intended by the researcher – when they have a moderate or neutral attitude to the issue in question but also for reasons of satisficing or social desirability. Nonetheless, most researchers recommend that a middle alternative should be offered in order to prevent respondents who have a moderate or neutral opinion from having to use an alternative category, thereby systematically distorting the data.

7. Non-substantive (DK) category

Two opposing positions have been adopted in the academic debate on the use of non-substantive – "don't know" (DK) or "no opinion" – response categories in rating scales. The classical position recommends that DK categories[1] should always be offered because it is assumed that respondents who do not have a relevant opinion on the issue in question would otherwise feel compelled to give a substantive answer. In other words, they would randomly choose a substantive response category instead of reporting the fact that they did not have a relevant opinion (e.g., Katz, 1942; Payne, 1950; Vaillancourt, 1973; Schuman & Presser, 1981; Converse & Presser, 1986). On the other hand, representatives of the more "modern" position argue that offering DK categories is problematic because they will be chosen not only by respondents who do not have a relevant opinion but also by satisficing respondents (e.g., Gilljam & Granberg, 1993; Krosnick & Fabrigar, 1997). Moreover, respondents might use a DK response to avoid expressing socially undesirable opinions or if they did not understand a question or had difficulties with the response alternatives (Krosnick & Fabrigar, 1997). Moreover, respondents might interpret the fact that a DK category is offered to mean that comprehensive knowledge is needed to answer the question, which could lead to uncertainty and thus to selection of the DK category (Hippler & Schwarz, 1989).

[1] In this section, the term *DK category* is used as a synonym for various non-substantive response categories such as *no opinion*, *don't know*, and *I can't say*.

An alternative to offering a DK category is to precede the question with a "DK filter" that asks whether the person has an opinion on the issue in question (Schuman & Presser, 1981). If this is the case, this opinion is asked about in detail. Otherwise, the next question is asked. The aim of DK filtering is (a) to avoid compelling respondents to give a substantively incorrect answer and (b) in so doing, to improve the quality of the data. However, a comparison of filtered and unfiltered questions revealed that the rate of DK responses was between 20 and 25% higher in filtered questions than in unfiltered questions (Schuman & Presser, 1981). The wording of the filter question has a considerable influence on whether respondents report that they do not have an opinion on the issue: When the filter question is worded more generally (e.g., "Do you have an opinion on this?"), respondents tend more to report that they have an opinion than when the filter question implies that it is necessary to have intensively engaged with the issue in order to be able to give an answer (e.g., "Have you thought/read enough about the issue to have an opinion on it?") (e.g., Bishop, Oldendick & Tuchfarber, 1983; Hippler & Schwarz, 1989, Krosnick & Abelson, 1991; Fowler & Cannell, 1996). The more abstract or unfamiliar the subject of the question is, the greater is the effect of filter wording (Bishop et al., 1983).

With regard to data quality, Andrews (1984) demonstrated that scales with a DK category achieved higher validity and lower method effects and error variances than scales without such a category. However, other experimental studies found that omitting DK categories did not influence data quality (e.g., Poe et al.,1988; Alwin & Krosnick, 1991; McClendon & Alwin, 1993; Krosnick et al., 2002). In an election study, more exact election forecasts were achieved when respondents who had chosen a DK response were subsequently pressed to give a substantive answer (Visser, Krosnick, Marquette, & Curtin, 2000). The content of the question and the degree of differentiation of the respondent's opinion may play a role in the selection of the DK category.

The form that a DK option takes depends on the survey mode. In postal or other paper-based surveys it must be decided whether or not to provide an explicit DK category. In face-to-face or telephone interviews, researchers can choose between explicitly offering a DK category or having the interviewer accept as a DK response an independently expressed report by the respondent that he or she does not have an opinion on the issue in question. Interviewers are frequently instructed to make one additional attempt to elicit a substantive response and, if this does not succeed, to record as a DK response the answer volunteered by the respondent. In interactive computer-assisted surveys there are various technical possibilities of implementing the DK option. A DK category is either explicitly offered or, if the respondent does not answer the question, he or she is immediately requested to give an answer or to confirm the DK response (implicit option). The two approaches can also be combined by explicitly offering a DK category and immediately asking a probing question if the respondent chose neither a substantive nor the DK category. DeRouvray and Couper (2002) found the lowest rate of item nonresponse in the case of a design in which an explicit DK category was not offered but an additional attempt was made to elicit a definitive answer from respondents who did not answer a question, thereby giving them the opportunity to confirm a DK response.

Conclusion and recommendation:

When deciding whether or not to provide a DK category, the question content, survey mode, and target group should be taken into account. For example, researchers must decide whether it might be problematic not to offer a DK category to a certain target group. If they are certain that the respondents know an answer, they can dispense with a DK option.

8. Likert-type agree-disagree and item-specific rating scales

Likert-type scales are rating scales in which the dimension is *agreement* – for example, agree/disagree or completely disagree/completely agree.

Likert-type scales have been universally applied to different statements in so-called item batteries. In the International Social Survey Programme (ISSP) 2012, for example, the rating scale "strongly agree," "agree," "neither agree nor disagree," "disagree," "strongly disagree" was used for the evaluation of the following statements (Terwey & Baltzer, 2013):

> One parent can bring up a child as well as two parents together.
>
> A same sex female couple can bring up a child as well as a male-female couple.
>
> A same sex male couple can bring up a child as well as a male-female couple.

One alternative here would be to ask "How well can one parent bring up a child?" In this case, the dimension would be evaluative, for example *very badly–very well*. This type of response option is referred to as "item-specific" (Saris, Revilla, Krosnick & Shaeffer, 2010) and the corresponding rating scales are known as item-specific rating scales.

A number of studies show that agree/disagree scales encourage acquiescence (Billiet & McClendon, 2000). Krosnick and Presser (2010) compiled findings that showed that acquiescence was also very likely in the case of true/false and yes/no options. For this reason, the use of item-specific scales is recommended, especially in the case of statements such as "I am often sad" and "Short waiting times at the doctor are important to me." In the case of such statements, it is easier to have respondents directly assess frequency or importance. Moreover, it also increases the quality of the measurement, as has been shown for various countries (including Germany) in the European Social Survey (ESS; Saris et al., 2010).

Conclusion and recommendation: Empirical findings suggest that it is better to use item-specific scales and to avoid agree/disagree scales as they elicit higher rates of agreement than item-specific rating scales.

9. Graphic representation of scales

Experimental studies have shown that graphical elements of rating scales can systematically influence response behaviour because respondents use not only verbal but also nonverbal, visual elements of the questionnaire when interpreting and answering questions (e.g., Smith, 1995; Christian & Dillman, 2004; Tourangeau, Couper, & Conrad, 2004; Tourangeau, Couper, & Conrad, 2007; Christian, Parsons, & Dillman, 2009; Toepoel & Couper, 2011).

In experimental studies, significantly different response distributions have been observed in the case of vertical and horizontal rating scales (e.g., Friedman & Friedman, 1994; Toepoel et al., 2009), although the direction of the effect was not consistent. However, a number of studies have shown that primacy effects occurred in vertical rating scales and that therefore horizontal scales should be used instead (Tourangeau et al., 2000).

One fundamental element of the visual design of rating scales is the scale midpoint, because respondents orient themselves towards it when interpreting the scale. In some designs, the conceptual and visual scale midpoints do not coincide, for example (a) when a DK category is added as a further

category and is not differentiated visually from the substantive options by a line or a space, or (b) when the scale categories are not equidistant. An experimental comparison of rating scales in which the conceptual and visual midpoints either coincided or did not coincide revealed significantly different response distributions (e.g., Tourangeau et al., 2004; Christian et al., 2009).

Tourangeau et al. (2004) found that when extreme response options were represented by shades of a different colour (shades of blue on the disagreement side of the scale and shades of red on the agreement side), respondents tended to avoid the disagreement side more than when the two ends of the scale were represented by shades of the same colour (e.g., shades of blue). However, these effects no longer occurred when the scale was verbally labelled.

Conclusion and recommendation: In general, we recommend that non-task-related graphical elements such as colours, shading, or symbols should be used with caution in rating scales because they may lead to undesirable effects. It is important that the graphical representation should reflect the symmetry of the scale and the equidistance of the response options. For example the non-substantive categories should be visually differentiated from the rest of the rating scale. And finally, rating scales should be horizontally oriented in order to reduce primacy effects.

References

Alwin, D. F., & Krosnick J. A. (1991). The reliability of survey attitude measurement: The influence of question and response attributes. *Sociological Methods and Research, 20*, 139-181.

Andrews, F. M. (1984). Construct validity and error components of survey measures: A structural equation approach. *Public Opinion Quarterly*, *48*, 409-448.

Billiet, J., & McClendon, M. J. (2000). Modeling acquiescence in measurement models for two balanced sets of items. *Structural Equation Modeling, 7 (4):* 608-628.

Bishop, G. F., Oldendick R. W., & Tuchfarber, A. J. (1983). Effects of filter questions in public opinion surveys. *Public Opinion Quarterly*, *47*, 528-546.

Bishop, G. F., Oldendick, R. W., Tuchfarber, A. J., & Bennett, S. E. (1980). Pseudo-opinions on public affairs. *Public Opinion Quarterly, 44*, 198-209.

Christian, L. M., & Dillman, D. A. (2004). The influence of graphical and symbolic language manipulations on responses to self-administered questions. *Public Opinion Quarterly, 68* (1), 57-80.

Christian, L. M., Parsons, N. L., & Dillmann, D. A. (2009). Designing scalar questions for web surveys. *Sociological Methods and Research, 37*, 393-425.

Clark, D. A. (1990). Verbal uncertainty expressions: A critical review of two decades of research. *Current Psychology: Research and Reviews, 9*(3), 203-235.

Converse, J. M., & Presser, S. (1986). *Survey questions*. Beverly Hills: Sage Publications, Inc.

Danner, D. (2016). Reliability – The precision of a measurement. *GESIS Survey Guidelines*. Mannheim, Germany: GESIS – Leibniz Institute for the Social Sciences. doi: 10.15465/gesis-sg_en_011

DeRouvray, C., & Couper, M. P. (2002). Designing a strategy for reducing “no opinion” responses in web-based surveys. *Social Science Computer Review, 20*, 3-9.

Dubois, B., & Burns, J. A. (1975). An analysis of the meaning of the question mark response category in attitude scales. *Educational and Psychological Measurement, 35*, 869-884.

Fowler, F. J., & Cannell, C. F. (1996). Using behavioral coding to identify cognitive problems with survey questions. In N. Schwarz & S. Sudman (Eds.), *Answering questions: Methodology for determining cognitive and communicative processes in survey research* (pp. 15–36). San Francisco: Jossey-Bass.

French-Lazovik, G., & Gibson, C. L. (1984). Effects of verbally labeled anchor points on the distributional parameters of rating measures. *Applied Psychological Measurement, 8*(1), 49-57.

Friedman, L. W., & Friedman, H. H. (1994). A comparison of vertical and horizontal rating scales. *The Mid-Atlantic Journal of Business, 30,* 107-202.

Gilljam, M., & Granberg, D. (1993). Should we take don't know for an answer? *Public Opinion Quarterly, 57*(3), 348-357.

Hippler, H. J., & Schwarz, N. (1989). "No opinion" filters: A cognitive perspective. *International Journal of Public Opinion Research, 1*, 77-87.

Hofmans, J., Theuns, P., Baekelandt, S., Mairesse, O. Schillewaert, N., & Cools, W. (2007). Bias and changes in perceived intensity of verbal qualifiers affected by scale orientation. *Survey Research Methods 1*, 97–108.

Kalton, G., Robert, J., & Holt, D. (1980). The effects of offering a middle response option with opinion questions. *Statistician, 29,* 65-78.

Kaplan, K. J. (1972). On the ambivalence-indifference problem in attitude theory and measurement. A suggested modification of the semantic differential technique. *Psychological Bulletin*, 77, 361-372.

Katz, D. (1942). Do interviewers bias pool results? *Public Opinion Quarterly, 6,* 248-268.

Krebs, D. (2012). The impact of response format on attitude measurement. In S. Salzborn, E. Davidov, & J. Reinecke (Eds.), *Methods, Theories, and Empirical Applications in the Social Sciences* (pp. 105-113). Wiesbaden: VS Verlag für Sozialwissenschaften.

Krebs, D., & Hoffmeyer-Zlotnik, J. H. P. (2010). Positive first or negative first? *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 6*(3), 118-127.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology, 5,* 213-236.

Krosnick, J. A. (2002). The causes of no-opinion responses to attitude measures in surveys: They are rarely what they appear to be. In R. M. Groves, ,Don A. Dillman, John N. Eltinge & R. J. A. Little (Eds.), *Survey Nonresponse* (pp. 88-100). New York: Wiley-Interscience.

Krosnick, J. A., & Abelson, R. P. (1991). The case for measuring attitude strength in surveys. In J. M. Tanur (Ed.), *Questions about survey questions (pp. 177-203)*. New York: Russell Sage.

Krosnick, J. A., & Alwin, D. F. (1987). An evaluation of a cognitive theory of response-order effects in survey measurement. *Public Opinion Quarterly, 51*, 201-219.

Krosnick, J. A., & Fabrigar L. R. (1997). Designing rating scales for effective measurement in surveys. In L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, & D. Trewin (Eds.), *Survey measurement and process quality* (pp. 141-164). New York: John Wiley & Sons, Inc.

Krosnick, J. A., Holbrook, A. L., Berent, M. K., Carson, R. T., Hanemann, W. M., Kopp, R. J., Mitchell, R. C., Presser, S., Rudd, P. A., Smith, V. K., Moody, W. R., Green, M. C., & Conaway, M. (2002). The impact of "no opinion" response options on data quality: non-attitude reduction or an invitation to satisfice? *The Public Opinion Quarterly, 66*, 371-403.

Krosnick, J. A., & Presser S. (2010). Question and Questionnaire Design. Peter V. Marsden and James D. Wright (eds.), *Handbook of Survey Research*, (pp. 264-313). Bingley, UK: Emerald.

Krosnick, J. A., Narayan, S. S., & Smith, W. R. (1996). Satisficing in surveys: Initial evidence. In M. T. Braverman & J. K. Slater (Eds.), *Advances in survey research* (pp. 29-44). San Francisco: Jossey-Bass.

Krosnick, J. A., & Presser, S. (2010). Question and questionnaire design. In J. D. Wright & P. V. Marsden (Eds.), *Handbook of survey research* (second edition) (pp. 263-313). Bingley, UK: Emerald Group.

Kulas, J. T., Stachowski, A. A., & Haynes, B. A. (2008). Middle response functioning in Likert-responses to personality items. *Journal of Business and Psychology, 22*(3), 251-259.

Lenzner, T., Neuert, C., & Otto, W. (2016). Cognitive Pretesting. *GESIS Survey Guidelines*. Mannheim, Germany: GESIS – Leibniz Institute for the Social Sciences. doi: 10.15465/gesis-sg_en_010

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

Maitland, A. (2009). How many scale points should I include for attitudinal questions? *Survey Practice 06*. AAPOR e-journal.

McClendon M. J., & Alwin, D. F. (1993). No-opinion filters and attitude measurement reliability. *Sociological Methods Research, 21*, 438-464.

Menold, N., & Bogner, K. (2012). Antwortskalen in sozialwissenschaftlichen Umfragen: Theoretische Modelle, Stand der Forschung und Forschungsperspektiven. In H.-G. Soeffner, (Ed.): *Transnationale Vergesellschaftungen. Verhandlungen des 35. Kongresses der Deutschen Gesellschaft für Soziologie in Frankfurt am Main 2010* (CD-ROM).Wiesbaden: VS Verlag.

Menold, N., Kaczmirek, L., Lenzner, T., & Neusar, A. (2014). How do respondents attend to verbal labels in rating scales? *Field Methods, 26*(1), 21-39.

Menold, N., & Kemper, Ch. J. (2015). The impact of frequency rating scale formats on the measurement of latent variables in web surveys – An experimental investigation using a measure of affectivity as an example. *Psihologija 48 (4), 431-449.* doi: http://dx.doi.org/10.2298/PSI1504431M.

Menold, N., & Raykov, T. (2015). Can reliability of multiple component measuring instruments depend on response option presentation mode? *Educational and Psychological Measurement,* online first. doi: http://dx.doi.org/10.1177/0013164415593602.

Menold, N., & Tausch, A. (2015). Measurement of latent variables with different rating scales: Testing reliability and measurement equivalence by varying the number of categories and verbalization. *Sociological Methods and Research.* doi: 10.1177/0049124115583913.

Moors, G., Kieruj, N., & Vermunt, J. K. (2014). The effect of labeling and numbering of response scales on the likelihood of response bias. *Sociological Methodology 44* (1), 369-399.

O'Muircheartaigh, C., Krosnick, J. A., & Helic, A. (1999). Middle alternatives, acquiescence, and the quality of questionnaire data. Paper presented at the *annual meeting of the American Association for Public Opinion Research*, St. Petersburg, Florida.

Pajares, F., Hartley, J., & Vahante, G. (2001): Response format in writing self-efficacy assessment: Greater discrimination increases prediction. *Measurement and Evaluation in Counseling and Development, 33,* 214-221.

Parducci, A. (1983). Category ratings and the relational character of judgment. In H. G. Geissler, H. F. J. M. Bulfart, E. L. H. Leeuwenberg & V. Sarris, *Modern Issues in Perception* (pp. 262-282). Berlin: VEB Deutscher Verlag der Wissenschaften.

Payne, S. L. (1950). Thoughts about meaningless questions. *Public Opinion Quarterly, 14*, 687-696.

Poe, G. S., Seeman, I., McLaughlin, J., Mehl, E., & Dietz, M. (1988). Don't know boxes in factual questions in a mail questionnaire. *Public Opinion Quarterly, 52*, 212-222.

Pollack, S., Friedman H. H., & Presby L. (1990). Two salient factors in the construction of rating scales: Strength and direction of anchoring adjectives. *International Conference of Measurement Errors in Surveys*, Tucson, Arizona, November 11-14, p. 57.

Preston, C. C., & Colman, A. (2000). Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica, 104*(1), 1-15.

Rohrmann, B. (1978). Empirische Studien zur Entwicklung von Antwortskalen für die sozialwissenschaftliche Forschung. *Zeitschrift für Sozialpsychologie*, *1978*, 222-245.

Saris, W. E., & Gallhofer, I. N. (2007). *Design, evaluation, and analysis of questionnaires for survey research*. Hoboken, New Jersey: John Wiley & Sons, Inc.

Saris, W., Revilla, M., Krosnick, J. A., & Shaeffer, E. (2010). Comparing questions with agree/disagree response options to questions with item-specific response options. S*urvey Research Methods, 4,* 61-79.

Schaeffer, N. C., & Barker, K. (1995). Issues in using bipolar response categories: Numeric labels and the middle category. Paper presented at the *annual meeting of the American Association for Public Opinion Research,* Ft. Lauderdale, FL, May 23, 1995.

Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys: Experiments on question form, wording and context.* New York: Academic Press.

Schwarz, N., Bless, H., Bohner, G., Harlacher, U., & Kellenbenz, M. (1991). Response scales as frames of reference: The impact of frequency range on diagnostic judgment. *Applied Cognitive Psychology*, 5, 37-50.

Smith, T. W. (1995). Little things matter: A sampler of how differences in questionnaire format can affect survey responses. *Proceedings of the American Statistical Association, Survey Research Methods Section*: 1046-1051.

Sturgis, P., Roberts, C., & Smith, P. (2014). Middle alternatives revisited: How the neither/nor response acts as a way of saying "I don't know"? *Sociological Methods & Research, 43*(1), 15-38.

Terwey, M., & Baltzer, S. (2013): Variable Report ALLBUS / Allgemeine Bevölkerungsumfrage der Sozialwissenschaften 2012. ZA-Nr. 4614. Köln: GESIS, GESIS - Variable Reports; No. 2013/16.

Theil, M. (2002). The role of translations of verbal into numerical probability expressions in risk management: a meta-analysis. *Journal of Risk Research, 5*(2), 177-186.

Thurstone, L. L. (1929). Theory of attitude measurement. *Psychological Review*, *36*(3), 222-241.

Toepoel, V. (2008). *A Closer Look at Web Questionnaire Design*. Tilburg: Tilburg University Press.

Toepoel, V. & Couper, M. P. (2011). Can verbal instructions counteract visual context effects in web surveys? *Public Opinion Quarterly*, *75*(1), 1-18.

Toepoel, V., Das, M., & van Soest, A. (2009). Design of web questionnaires: The effect of layout in rating scales. *Journal of Official Statistics, 25*, 509-528.

Toepoel, V., & Dillman, D. A. (2011). Words, numbers, and visual heuristics in web surveys: Is there a hierarchy of importance? *Social Science Computer Review, 29*(2), 193-207.

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

Tourangeau, R., Couper, M. P., & Conrad, F. G. (2007). Color, labels, and interpretive heuristics for response scales. *Public Opinion Quarterly, 71,* 91-112.

Tourangeau, R., Rips, L. J., & Rasinski, K. A. (2000). *The Psychology of Survey Response*. Cambridge: Cambridge University Press.

Vaillancourt, P. M. (1973). Stability of children's survey responses. *Public Opinion Quarterly*, 37(3), 373-387.

Visser, P. S., Krosnick, J. A., Marquette, J. F., & Curtin, M. F. (2000): Improving election forecasting: Allocation of undecided respondents, identification of likely voters, and response order effects. In P. Lavrakas & M. W. Traugott (Eds.), *Election polls, the news media, and democracy*. New York: Chatham House.

Wallsten, T. S., Budescu, D. V., & Zwick, R. (1993). Comparing the calibration and coherence of numerical and verbal probability judgments. *Management Science, 39,* 176-190.

Windschitl, P. D., & Wells, G. L. (1996). Measuring psychological uncertainty: Verbal versus numeric methods. *Journal of Experimental Psychology: Applied, 2*(4), 343-364.

Worcester, R. M., & Burns, T. R. (1975). A Statistical Examination of the Relative Precision of Verbal Scales. *Journal of the Market Research Society 17* (3), 181–197.

Zaller, J. R. (1988). Vague questions vs. vague minds: Experimental attempts to reduce measurement error. Paper presented at the *annual meeting of the American Political Science Association*, Washington, D.C.