Rating-Scale Labeling in Online Surveys: An Experimental Comparison of Verbal and Numeric Rating Scales with Respect to Measurement Quality and Respondents' Cognitive Processes

Menold, Natalja

Sociological Methods & Research
2020, Vol. 49(1) 79-107
© The Author(s) 2017
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0049124117729694
journals.sagepub.com/home/smr
SAGE

Rating-Scale Labeling in Online Surveys: An Experimental Comparison of Verbal and Numeric Rating Scales with Respect to Measurement Quality and Respondents' Cognitive Processes

Natalja Menold[1]

Abstract

Unlike other data collection modes, the effect of labeling rating scales on reliability and validity, as relevant aspects of measurement quality, has seldom been addressed in online surveys. In this study, verbal and numeric rating scales were compared in split-ballot online survey experiments. In the first experiment, respondents' cognitive processes were observed by means of eye tracking, that is, determining the respondent's fixations in different areas of the screen. In the remaining experiments, data for reliability and validity analysis were collected from a German adult sample. The results show that respondents needed more fixations and more time to endorse a category when a rating scale had numeric labels. Cross-sectional reliability was lower

[1] GESIS—Leibniz Institute for the Social Sciences, Mannheim, Germany

Corresponding Author:
Natalja Menold, GESIS—Leibniz Institute for the Social Sciences, P.O. Box 12 21 55, D-68072 Mannheim, 68159, Germany.
Email: natalja.menold@gesis.org

and some hypotheses with respect to the criterion validity could not be supported when numeric rating scales were used. In conclusion, theoretical considerations and the empirical results contradict the current broad usage of numeric scales in online surveys.

Keywords

rating scales, labeling, eye tracking, composite reliability, criterion validity

Questions and items with rating scales are commonly used in social sciences, where online surveys are becoming increasingly popular. A rating scale is composed of a measurement continuum that extends from one extreme to the other (e.g., strongly agree–strongly disagree). When developing rating scales, researchers are faced with the challenge of making numerous decisions, for example, regarding the number of categories, category labeling, and the choice between unipolar and bipolar scales or deciding about the order of categories.

This article focuses on a comparison between verbal and numeric rating scales, because labeling has been recognized as a basic cue that influences the clarity of the rating-scale categories and their understanding (Parducci 1983). In verbal rating scales, a verbal label (e.g., strongly agree, rather agree, rather disagree, strongly disagree) is used for each category, while, in numeric rating scales, verbal labels are used to mark the end categories only, and numbers (e.g., 0–4) are used to additionally mark all categories.

Large-scale population surveys, in general, use both verbal and numeric rating scales, but standards with respect to verbalization have not yet been established. Examples include the European Value Study[1] (a face-to-face survey), the European Social Survey[2] (a face-to-face survey), the German Longitudinal Election Study[3] (GLES, a multi-mode survey), the GESIS-Panel (a probability sample online panel),[4] and the LISS (a Dutch probability sample online panel).[5] Some surveys tend to prefer using verbal rating scales, as, for instance, the German General Social Survey (GGSS/ALLBUS).[6] Furthermore, surveys usually use either verbal or numeric rating scales in a rating scale, whereas the use of both verbal and numeric labels seems to be relatively uncommon. This information on the usage of verbal and numeric labels was extracted from the questionnaires or from show card files. It is important to note that examples from psychological and educational research have not been considered, because the article focuses on social science research.

The described broad usage of numeric rating scales contradicts theoretical considerations regarding their possible psychological and cognitive effects. According to Windschitl and Wells (1996), two systems of reasoning have been differentiated in psychology and philosophy (e.g., by Epstein 1990 or by Vygotsky 1934). The first is more rational, deliberative, and rule-based, whereas the second is rather associative, spontaneous, and intuitive. At a given point in time, one system can dominate over the other, which has implications for individual reactions and responses. When measuring attitudes in surveys, respondents are actually supposed to react spontaneously, because it is less natural to express one's opinions and attitudes very exactly. Therefore, according to Windschitl and Wells (1996), a survey context is more compatible with the associative reasoning system, in which the usage of verbal rating scales suits the given context. By contrast, numeric scales are compatible with rational reasoning and their usage for (attitude) measurements in surveys might lead to a conflict between the situational context and the dominating reasoning system, which could result in lower data quality.

Krosnick and Fabrigar (1997) also assume that numeric rating scales are less natural and, therefore, more difficult for respondents to use for their self-descriptions. They additionally assume that numbers might not have an inherent meaning for the respondents besides providing a rank order of categories. Krosnick and Fabrigar (1997) therefore stipulate that verbal rating scales are associated with less satisficing than are numeric rating scales. Satisficing by respondents is defined by Krosnick and Alwin (1987) as a low road of information processing that leads to low data quality. For example, respondents do not make any effort to access their responses or do not attend to all the information that needs to be considered. The tendency to satisfice is particularly high in the case of low motivation, reduced cognitive abilities, and high task difficulty. Therefore, survey tasks should be as clear and easy as possible to evoke the fewest possible biased reactions from respondents.

This article focuses on measurement quality of rating scales in online surveys, which is evaluated by means of reliability and validity metrics. Reliability and validity are very relevant concepts when assessing data quality (Groves et al. 2004). Reliability describes measurement precision and is defined as a ratio of true variance to observed variance (Lord and Novick 1968). Validity is defined as the extent to which a result accurately points to the content (construct) of interest and is measured as a function of the correlation between the survey statistic and the true value of the construct (Groves et al. 2004:25). The main goal of the measurement is to generalize from the data to a certain concept and to provide conclusions with respect to this concept. Assessing validity (construct or criterion validity; see e.g., Rammstedt et al. 2015) is one way of ensuring that the generalization is acceptable. A high validity can be only achieved when reliability is high or sufficient (e.g., Raykov and Marcoulides 2011). Therefore, reliability and validity are more direct and central metrics of measurement quality and should be addressed and documented for every measurement instrument (Rammstedt et al. 2015).

Research on comparisons of reliability between verbal and not fully verbalized forms in non-online mode has been conducted for decades. Some studies compared verbal forms with the forms including verbalized extreme categories without any other labels for intermediate categories (referred to as END form). These studies found a higher reliability for the verbal form (Alwin and Krosnick 1991; Menold et al. 2014; Weng 2004). Other studies, which did not explicitly differentiate between the numeric and END forms, obtained mixed results. While Saris and Gallhofer (2007) reported higher reliability for the verbal forms, Churchill and Peter (1984) did not find any differences and Andrews (1984) reported higher reliability for not fully verbalized forms. However, the results of these studies could not be clearly evaluated as relevant for the comparison of verbal forms with the END forms, or with the numeric forms, or both.

As far as the comparison between verbal and numeric rating scales in nononline mode is concerned, Krosnick and Berent (1993) found greater retest reliability with verbal than with numeric rating scales (only seven categories were tested). This was demonstrated by numerous experiments conducted as telephone, paper-and-pencil, and face-to-face surveys with student and general population samples. Higher interrater reliability for seven-category verbal than for seven-category numeric rating scales was also found by Peters and McCormick (1966). However, Finn (1972) did not observe any effects on reliability when using five categories. When evaluating the reported results, one should keep in mind that the studies by Peters and McCormick (1966) and Finn (1972) rely on very small student samples (overall $n < 40$). The studies by Krosnick and Berent (1993), Peters and McCormick (1966), and Finn (1972) did not use agree–disagree rating scales but requested evaluations or estimations of amounts.

Menold and Tausch (2016) found that a five-category agree–disagree numeric rating scale provided the lowest cross-sectional reliability, which was compared with application dimension (does not apply at all–applies fully) when using the END and verbal forms with five and seven categories. This study was conducted with students in a paper-and-pencil mode and focused on the comparison between the verbal and END forms. Here, only five- and seven-category rating scales were compared, because five to seven categories were found to be associated with the highest measurement quality in a number of previous studies (Krosnick and Fabrigar 1997). Menold and Tausch (2016) used items on the opinions about the European Union (EU), taken from the GLES. To obtain a benchmark to the GLES data, the authors realized an experimental group with the numeric rating scale from the GLES. A seven-category numeric rating scale was not incorporated into this study, due to its different focus.

With respect to validity of the non-online mode, some studies again report higher measurement quality for verbal than for numeric rating scales. Windschitl and Wells (1996), as well as Krosnick and Berent (1993), found higher criterion validity for verbal rating scales. Criterion validity provides support for hypotheses about the predictions of the values of an external variable (criterion) by examining the measure of interest. The result pertaining to the higher criterion validity of verbal rating scales, compared to numeric scales, was obtained for different modes, five (Windschitl and Wells 1996) and seven categories (Krosnick and Berent 1993), and different evaluation dimensions of the rating scales (likelihoods and amounts). Bendig and Hughes (1953), as well as Bernardin et al. (1976), reported greater differentiation with increased verbalization (for student samples and paper-and-pencil mode); this was observed for different numbers of categories and different types of ratings, such as amounts (know a lot–know a little) or desirability of events. However, in these both studies, only partly labelled rating scales were used. Finally, in a paper-and-pencil study with students, Newstead and Arnold (1989) did not observe a higher accuracy for a verbal frequency rating scale than for its numeric alternative.

To sum up the results for the non-online mode, evidence for a higher reliability and validity of verbal than numeric rating scales has been detected in some studies. For online surveys, only a few studies—those by Menold et al. (2014) and Menold and Kemper (2015)—on measurement quality, in terms of reliability and validity, are available. Of these, Menold et al. (2014) report higher cross-sectional reliability for verbal than for numeric rating scales, tested for five categories in a small sample of German adults. Menold and Kemper (2015) conducted an experimental survey by combining the number of categories (five vs. seven), verbalization (full vs. end), and usage of numbers in a nonprobability online panel in Germany. They used a psychological concept and a frequency rating scale. Particularly poor measurement quality, in terms of factorial validity and reliability, was obtained for the numeric rating scales and rating scales that combined both verbal and numeric labels. However, with a five-category verbal rating scale, unacceptably low reliability and factorial validity were obtained, whereas this was not the case with the seven-category verbal form.

Other studies that were conducted in online mode did not address reliability or validity of measurements. Nevertheless, their results generally favor verbal over numeric rating scales. Some studies (Toepoel and Dillman 2011; Tourangeau, Couper, and Conrad 2007) investigated the effect of the visual context (color shading of response options; uneven spacing) in probability and nonprobability online samples and found that a full verbalization of rating scales eliminated the context effects, which were given in rating scales with numeric labels and END forms. This was demonstrated for different numbers of categories (five or seven) and different types of ratings (e.g., favor–oppose, satisfaction, frequency). Results in favor of numeric rating scales were reported by Wouters et al. (2014) who found a higher variability for a numeric frequency rating scale, when compared with a verbal realization. However, the verbal rating scale in this study seems to not provide distances that appear to be equal, because it combines vague and more exact qualifiers (never, 1 time, several times, regularly, often); this could be an alternative explanation of the obtained result.

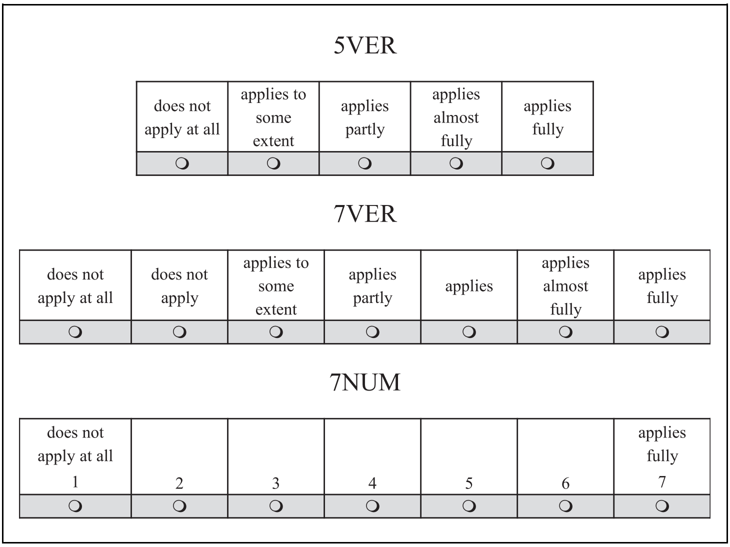
Because theoretical considerations and the majority of empirical results show that measurement quality is higher with verbal than with numeric forms, the widespread usage of the latter in the social science surveys is somewhat surprising. In the case of online surveys, this can be explained by the scarcity of evidence related to the measurement quality of numeric rating scales in terms of reliability and validity in this data collection mode. Differences in the effects of rating scale forms—between visual presentation on the paper (paper-and-pencil mode, show cards in face-to-face surveys) and online surveys—may be due to the preferences of respondents for reading electronic or printed texts or to differences in meta-cognitive processes (e.g., Ackerman and Goldsmith 2011).

In addition, relatively little is known about the differences in the underlying cognitive mechanisms and information processing when using verbal versus numeric rating scales. Knowledge about such processes is relevant for a better understanding of why and how rating-scale forms affect response behavior and, therefore, why an effect on measurement quality can be expected. Therefore, studies that address the cognitive process of respondents will help to correct or enrich the corresponding theories.

As far as the cognitive process is concerned, some of the empirical results available to date tend to contradict Krosnick and Fabrigar's (1997) assumption that more satisficing occurs with numeric rating scales than with verbal ones. Tourangeau et al. (2007) found longer response times for verbal than for numeric rating scales and explained this result as a consequence of the fact that respondents simply needed more time to read verbal labels, which hints at a higher cognitive effort in the case of verbal labels but not necessary at more satisficing. Menold et al. (2014), using eye tracking, also found that respondents needed more fixations and longer fixation times to attend to verbal than to numeric rating scales. The authors additionally observed that respondents did not attend to all verbal labels when providing their responses, which indicates the presence of satisficing behavior. In the study of Menold et al. (2014), a comparison of the numeric and verbal rating scales was conducted for the five categories only, so that no information is available for the respondents' cognitive processing and attention in the case of seven-category numeric rating scales. Because Menold and Tausch (2016) and Menold and Kemper (2015) found that the effect of verbalization on the measurement quality was different for seven and five categories, there is a need for studies on the cognitive response process associated with the usage of numeric and verbal labels in surveys that focus particularly on seven-category rating scales.

The present research was designed to continue the research reported by Menold et al. (2014) and Menold and Tausch (2016), with a focus on the cognitive process and measurement quality associated with the usage of verbal and numeric labels in seven-category rating scales. In addition, information on reliability and validity of verbal and numeric rating scales is obtained from a large online sample, which was not the case previously, because Menold and Tausch (2016) used a student sample and paper-and-pencil mode and the study by Menold et al. (2014) was a small sample study. The present research uses measurements of opinions about the EU, a concept, and some items that were also used by Menold et al. (2014) and Menold and Tausch (2016), to ensure comparability of results between the studies. It includes two additional concepts to increase generalization possibilities among different concepts. In contrast to the previous research in online mode, which used agreement (Menold et al. 2014) and frequency dimensions of rating scales (Menold and Kemper 2015), an application dimension (Figure 1) is used in the present study, similar to the study of Menold and Tausch (2016). The present research therefore provides new insights with regard to the cognitive process involved in using verbal and numeric rating scales. In particular, it contributes to the literature on online surveys through its focus on measurement quality in terms of reliability and validity.

The following hypotheses concerning differences between verbal and numeric rating scales, which were formulated taking into account the research results presented above, were tested:



5VER

does not apply at all	applies to some extent	applies partly	applies almost fully	applies fully
○	○	○	○	○

7VER

does not apply at all	does not apply	applies to some extent	applies partly	applies	applies almost fully	applies fully
○	○	○	○	○	○	○

7NUM

does not apply at all 1	2	3	4	5	6	applies fully 7
○	○	○	○	○	○	○

Figure 1. Rating scales in experimental groups.

Hypothesis 1 (H1): Dealing with numeric and verbal rating scales differs in terms of the underlying cognitive process.

Hypothesis 2 (H2): Numeric labels are associated with lower reliabilities than are verbal labels.

Hypothesis 3 (H3): Numeric labels are associated with lower validity than are verbal labels.

Methods and Data

Eye Tracking Study

In this article, the results of two studies are presented. The first study addressed respondents' cognitive process (H1) in an eye tracking experiment. It was conducted as an online survey in the laboratory during October and November 2012 at GESIS—Leibniz Institute for the Social Sciences, Germany. The procedure is described by Lenzner, Kaczmirek, and Galesic (2014). The present eye tracking experiment included evaluation of ten items and was a part of a large laboratory study that lasted longer than one hour. The experiments in the entire study were not related to each other and randomized assignment to an experimental group was conducted before each experiment, to avoid systematic presentation effects. For their participation in the entire study, respondents received a compensation of 30 Euros. A heterogeneous adult sample ($N = 82$) was used. The mean age of respondents was 36.2 years ($SD = 14.42$); 53.7 percent were females, and 68.3 percent had at least 12 years of schooling (German *Abitur*). The native language of 94 percent of respondents was German (the language in which the survey was conducted), 88 percent used the Internet daily. There was no difference between the randomized experimental groups (see below) with respect to all these background variables (lowest $p = .21$).

The apparatus for eye tracking used in the study is described by Lenzner et al. (2014:32). To record the participants' eye movements, the Tobii T120 eye tracking system was used; for the data analysis, Tobii Studio 3.2.1 software was used. T120 is specified to be accurate within 0.5°, with less than 0.3° drift over time, and less than 1° due to head motion. It allows for head movement within a 30 × 22 × 30 cm volume centered up to 70 cm from the camera. The sampling rate was 120 Hz (120 data points per second were collected for each eye), which allows for an unequivocal determination that a particular point on the screen was fixed by respondent's eyes. Respondents' baseline fixation length and baseline fixation count were measured at the end of the eye tracking session with two questions about the German government (see Lenzner et al. 2014, for further details). Baseline variables were included in the analyses as a covariate variable, to control for differences in individual respondents' reading characteristics.

The study was conducted to enrich the work of Menold et al. (2014) and Menold and Tausch (2016). The construct under consideration was opinions on the EU, used also in both these studies, which was measured in the present study with ten items from the GLES 2011 (Rattinger et al. 2011). In the eye tracking study, the ten EU items were presented on the two online survey pages.

Menold et al. (2014) found a two-factorial structure of the EU items from GLES. The first factor (items 1, 2, 3, 5, 6, 7; see Appendix) contained items that describe relatively negative aspects of the EU's impact on German life and the economy and the preference for retaining state sovereignty within the EU. The second factor covered the remaining items that state the positive impact of the EU on German life and the economy and a positive evaluation of the Euro. The author of the present article also conducted a confirmatory factor analysis (CFA; software Mplus 7.0, robust maximum likelihood (MLR) estimator) with the GLES 2014 data (Rattinger et al. 2014) to confirm the two-dimensional structure of the items used in the present research.

The two-factor model yielded reasonable goodness-of-fit statistics according to Root Mean Square Error of Approximation (RMSEA) and Comparative Fit Index (CFI), $\chi^2_{(30,\, N\,=\,1{,}010)} = 100.57$, $p = .000$, RMSEA $= 0.05$ (90 percent CI: [0.04, 0.06]), CFI $= .94$; $r_{(\text{between the factors})} = -.73$; standardized loadings ranged from $\lambda = 0.33$ to $\lambda = .75$ for the first factor and from $\lambda = .40$ to $\lambda = .67$ for the second factor. However, three covaried errors (between items 5 and 7, which both address the eastward expansion of the EU; between items 4 and 5, which both address economic consequences; and between items 9 and 3, which both address the EU regulations) as well as a small cross-loading $\lambda = .30$ of the item 10 on the first factor had to be modeled.

GLES used a five-category agree–disagree numeric rating scale with the EU items. The results demonstrating its poor reliability are available from the studies of Menold et al. (2014) and Menold and Tausch (2016), although the latter was not an online survey. Therefore, the five-category numeric form was no longer used and the five-category verbal form (5VER) was included in the present study, to obtain a benchmark for the studies by Menold et al. (2014) and Menold and Tausch (2016).

In the present eye tracking experiment, three rating-scale versions were implemented: a five-category verbal form (5VER), a seven-category verbal form (7VER), and a seven-category numeric form (7NUM), with numbers from 1 to 7 for each category. The rating-scale forms used in the study are shown in Figure 1.

The author included these three rating-scale groups in order to obtain a reasonable number of cases in each group, because it was not possible to include more participants in the eye tracking study, due to financial restrictions. The sample size of $N = 82$ is small, but large enough to uncover significance for mean differences of the eye tracking metrics between the three groups (the lowest $n = 25$; see Table 1) in the case of a large effect size and high practical relevance (Cohen 1988).

Validation Study

The second study was conducted in September 2014 at GESIS, Germany. Panelists of an online panel ($N = 497$) participated in the study. The online panel covered a heterogeneous sample of German adults (internet users). Of the study participants, 55 percent was male and 50 percent had completed senior high school (university entrance diploma or German *Abitur*). The mean age was 48 years ($SD = 15.72$). None of the experimental groups (see below) significantly differed with respect to these respondents' variables ($p > .10$), except gender, $\chi^2_{(2,\, N\,=\,486)} = 6.01$, $p = .05$, in experiment 2

Table 1. Fixation Times on Different Areas of the Screen in the Three Rating Scale Groups.

Rating Scale	5VER		7VER		7NUM			
Screen areas	Mean	*SD*	Mean	*SD*	Mean	*SD*	*F*(2, 79)	Part. η^2
Introduction	4.13	3.73	3.16	2.05	4.06	2.61	1.23	.03
Items	32.66	17.81	33.95	24.33	29.38	13.41	0.45	.01
Rating scale	13.94	6.61	10.21	5.94	8.13	4.59	5.57**	.13
Answer field	13.91	5.35	15.48	8.57	21.77	11.47	8.19***	.17
Total	67.00	28.37	65.17	35.11	65.70	21.75	.53	.01
n	25		28		29			

Note: MANCOVA results with respect to the main effects of (1) rating scale groups: Wilks'λ = 0.63; $F_{(10,\ 148)}$= 3.80; p = .00; η^2 = 0.20; (2) respondents' baseline fixation time (covariate): Wilks'λ = 0.60; $F_{(5,\ 74)}$= 9.77; p = .00; η^2 = 0.40. MANCOVA = multivariate analysis of covariance; 5VER = five-category verbal form; 7VER = seven-category verbal form; 7NUM = seven-category numeric form; *SD* = standard deviation.
***p* < .01. ****p* < .001.

(see below), where there were 49 percent males in the 5VER group, 62 percent males in the 7VER group, and 55 percent males in the 7NUM group.

Three randomized experiments—with a total duration of about five minutes (time needed to respond to the experimental stimuli)—were included in the online survey to test H2 and H3. The whole online survey took approximately 20 minutes and included additional questions, for example, on values, authoritarianism, and demographic characteristics; their presentation forms did not vary among experimental groups and they were used, in part, to obtain validity measures for the present experiments. The participants received 5 Euros as an incentive for their participation in the entire survey.

The first experiment of the validation study was exactly the same as the one for the eye tracking study. The ten EU items (Appendix) were used with the three rating-scale forms: five-category verbal form (5VER), seven-category verbal form (7VER), and seven-category numeric form (7NUM; Figure 1). In the second and third experiments, two concepts on gender roles were incorporated: the first describes consequences of women's labor force participation for parenting and the second concerns general ideologies about gender role (Braun 2006). The six items on gender roles were taken from the ALLBUS 2008, which were found, by the author's own analyses of the ALLBUS 2008 data, to represent two dimensions (CFA with Mplus, MLR: $\chi^2_{(8,\ N\,=\,3{,}447)} = 56.52, p < .001$; RMSEA = .04; CFI = .99, $r_{(\text{between the two factors})} = -.81$; loadings of items ranged from $\lambda = .54$ to $\lambda = .88$). The first dimension with three items can be interpreted as general gender role ideologies (further referred to as gender ideologies factor) and the second dimension as consequences of women's labor force participation for parenting (further referred to as parenting factor). The six ALLBUS 2008 items used in the present experiments are shown in Appendix.

The three parenting factor items were used in the second experiment of the validation study. In this experiment, the rating-scale groups from the first experiment were again realized. However, for the CFA, which is the basis of the reliability analysis (e.g., Raykov and Marcoulides 2011), large samples (or subsamples) are of advantage. Therefore, in experiment 3, only two experimental groups with seven-category verbal and numeric forms (7VER and 7NUM) were realized, because the study focuses on seven-category rating scales. This made it possible to obtain larger subsamples than in the other two experiments. In the third experiment, gender ideologies items from the battery on gender roles from the ALLBUS 2008 were used (see Appendix).

The reliability of multi-item measures can be accessed through cross-sectional reliability metrics. In previous research on rating scales, Cronbach's α has (Cronbach 1951) been predominantly used (e.g., Churchill and Peter 1984). However, α can be interpreted in terms of reliability only if items measure a factor with equal loadings and if there are no covaried error terms between the items (τ-equivalent measures). Otherwise, α not only overestimates but also underestimates reliability (e.g., Raykov 2001). As the assumptions of τ equivalence are often violated in social science measurements, less restrictive reliability methods have been recommended (e.g., Raykov and Marcoulides 2011; Schweizer 2011). In this article, we used factor-analysis-based estimation of reliability according to latent variable modeling (McDonald 1999; Raykov and Marcoulides 2011:160–161). This composite reliability method differs markedly from the Cronbach's α coefficient and relies on the assumption that items represent a measure of the latent dimension, whereby other assumptions are relaxed (congeneric model). Within the congeneric model, the decomposition into the true and error scores for a single measure is described as

$$X_i = b_i T + E_i, \tag{1}$$

where $X_1, \ldots, X_p$ are values of the observed measures ($i = 1, \ldots, p$), T is its underlying true score, and E_i is the error score. The composite reliability ($\hat{\rho}_x$ is measured within this framework as follows (Raykov and Marcoulides 2011:161):

$$\widehat{\rho_x} = \frac{(\hat{b}_1 + \cdots + \hat{b}_p)^2}{(\hat{b}_1 + \cdots + \hat{b}_p)^2 + \hat{\theta}_1 + \cdots + \hat{\theta}_p}, \tag{2}$$

where $b_1, \ldots, \ b_p$ are the factor loadings and $\theta_1 \ldots, \theta_p$ are the error variances obtained from one unidimensional CFA. This approach for accessing reliability is also used if a measuring instrument consists of more than one dimension referred to as general structure (Raykov 2012). When the general structure model is applied (Raykov 2012:485), items representing different dimensions of a concept are included in the preliminary CFA and the correlation between the factors is considered in equation (2). For the two dimensions of the EU items, an estimator for general structure was used, while, for both gender-role factors, equation (2) was used to assess reliability. To run the underlying CFAs, the Mplus software and the function robust to nonnormality (MLR) were used. In addition, an interval estimation of reliability, based on standard error (*SE*) estimations of Mplus, was conducted.

One option for empirically assessing validity is to obtain criterion validity through the prediction of the values of a variable (criterion) when using the values of another variable (predictor) whereby this prediction (or relationship) is derived from a theory and/or replicates well-known empirical findings (e.g., Rammstedt et al. 2015). The survey questions that are supposed to measure the criterion can be answered in the same survey (at the same point of time) as the questions for the given construct. This kind of criterion validity is often referred to as "concurrent validity." In the present study, concurrent criterion validity was analyzed in a fashion similar to that used in the studies of Windschitl and Wells (1996) and Krosnick and Berent (1993). Because these studies report strong validity differences between verbal and numeric rating scales in non-online mode and are based on a sufficiently large data base, comparability with these studies was important.

The following measures of external criteria were used to obtain the validity metrics: (1) authoritarianism measured by the short scale of authoritarianism (KSA-3; Beierlein et al. 2014) and (2) left–right self-placement measured by the question taken from the ALLBUS (Breyer 2015). Authoritarianism is a central ideological opinion and expresses motives for collective insurance, which can be reached at the cost of individual autonomy (Beierlein et al. 2014). Authoritarianism is described by conventional opinions against authorities, support for subsidies in the case of nonconformity, and rigid support for traditions and established norms (Altemeyer 1981). Because authoritarianism is expected to be positively related to negative political positions and negatively to positive political positions (e.g., Altemeyer 1981; Aichholzer and Zeglovits 2015), it can be used to obtain criterion validity for both EU opinions and gender roles. Support for the EU and for women's labor force participation is associated with support for nontraditional political development and the establishment of new political regulations and norms. Thus, a positive evaluation of the EU and nonconservative gender role opinions can be expected to be negatively related to authoritarianism. Corresponding relationships have been confirmed in the literature: Aichholzer and Zeglovits (2015) report a negative correlation between authoritarianism and support for the EU. In addition, Duncan, Peterson, and Winter (1997) found a positive relationship between the high scores of authoritarianism and traditional opinions with respect to gender roles.

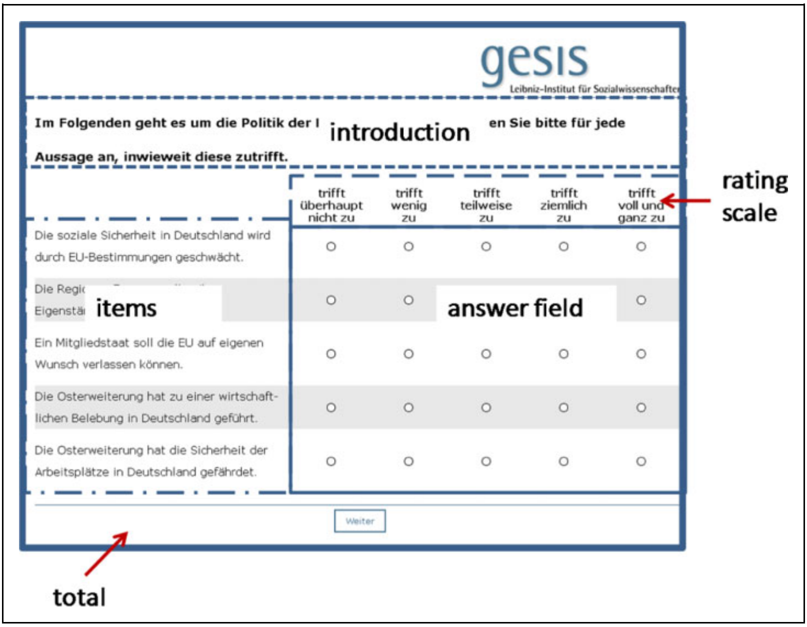
Left–right self-placement refers to the two central political orientations. In Germany, a right political orientation is associated with a more hierarchical and conservative attitude, while a left orientation refers to socialistic, progressive, and international orientation (Breyer 2015:3). Therefore, comparable with authoritarianism, one can expect that support for the EU, and women's labor force participation will be more strongly associated with left political orientation, whereas criticism of the EU and a negative view of women's labor force participation will be more closely associated with right political orientation.

To evaluate the criterion validity, scores for both EU factors and the two factors of gender roles (parenting and gender ideologies) were calculated. Next, a score for authoritarianism was obtained (high values reflecting high authoritarianism). The scores were mean values of the single item values calculated for each study participant. The values of item 2 of the parenting factor were recoded, so that high scores on the parenting factor reflect a positive evaluation of women's labor force participation. By way of contrast, the high scores of gender ideologies reflect a negative evaluation of women's occupational activity for family life. Similar to gender roles, the high scores for the EU factor 1 express negative impact of the EU on German life and the economy, while the high scores of the EU factor 2 reflect a positive evaluation of the EU. The left–right self-placement was measured on the numeric rating scale, which ranged from 1 (*left*) to 10 (*right*).

Results

Results of the Eye Tracking Study

With the eye tracking data, respondents' fixations on different screen areas were compared between the three experimental groups (5VER, 7VER, and



Im Folgenden geht es um die Politik der I en Sie bitte für jede Aussage an, inwieweit diese zutrifft.

	trifft überhaupt nicht zu	trifft wenig zu	trifft teilweise zu	trifft ziemlich zu	trifft voll und ganz zu
Die soziale Sicherheit in Deutschland wird durch EU-Bestimmungen geschwächt.	○	○	○	○	○
Die Regi Eigenstä	○	○			○
Ein Mitgliedstaat soll die EU auf eigenen Wunsch verlassen können.	○	○	○	○	○
Die Osterweiterung hat zu einer wirtschaftlichen Belebung in Deutschland geführt.	○	○	○	○	○
Die Osterweiterung hat die Sicherheit der Arbeitsplätze in Deutschland gefährdet.	○	○	○	○	○

Weiter

Figure 2. Presentation of the European Union (EU) items (first page) in the group five-category verbal form (5VER) with screen areas defined for the eye tracking study.

7NUM). The screen areas were defined to locate one single element of the online questionnaire page (Figure 2). First, an introductory text was presented, to obtain fixations needed to read the actual introduction. Second, the items' area was defined, to obtain fixations needed to read the items; third, the area including the rating scale was presented, to obtain fixations on rating-scale labels, and fourth, the area of the answer field, which is composed of radio buttons for respondents' answers, was defined. Finally, the total screen area was defined, which includes all the areas listed above as well as other areas not covered by these predefined screen areas. With the fixations on the rating-scale and answer areas, the respondents' cognitive process in the mapping stage of the cognitive response process (Tourangeau, Rips, and Rasinski 2000) can be observed. Fixations of the respondents were analyzed with the help of two variables: fixation time, which measures the time needed for all fixations to a predefined screen area, and fixation count, which measures the frequency of fixations on a screen area.

Differences in fixation times and fixation counts between the three rating-scale groups were obtained and tested for significance by means of a multivariate analysis of covariance with the SPSS 22 software. As described in the methods section, the respondents' baseline fixation time and fixation count were used as covariate variables (when addressing fixation times and fixation counts, respectively). Table 1 provides an overview of mean differences in fixation times between the rating-scale groups 5VER, 7VER, and 7NUM, as well as the results of the significance test. Table 1 shows that there were no significant differences between the rating-scale groups in fixations on the introduction and the items. Fixation times on the area of the rating scale were longer in both groups with verbal rating scales (5VER and 7VER) than in the 7NUM group. Interestingly, fixation times were highest for the 5VER rating scale, meaning that respondents did not need more time to fixate seven categories than five categories. Post hoc tests with Bonferroni correction (which was also used for all post hoc tests reported below) revealed a significant difference (*d*) in the fixation times on the rating-scale area between the 5VER group and other groups ($d_{5\text{VER/7VER}} = 3.92$, $p = .05$; $d_{5\text{VER/7NUM}} = 5.85$, $p = .001$), but not between the 7VER and 7NUM groups ($d = 1.92$ $p > .10$).

Next, fixation time on the answer field area was, with approx. 22 seconds, significantly longer in the 7NUM group than in the other two groups (Table 1). Post hoc tests showed that fixation time in the case of both 5VER and 7VER rating scales differed significantly from the fixation time in the 7NUM rating-scale group ($d_{(5\text{VER/7NUM})} = 9.08, p = .001$ and $d_{(7\text{VER/7NUM})} = 6.85$, $p = .01$), whereas the difference between 5VER and 7VER rating-scale groups ($d = 2.24$) was not significant ($p = 1.00$). In spite of the differences obtained for the rating-scale and answer areas, total fixation time did not significantly differ between the rating-scale groups (Table 1).

The results for the fixation count were strongly comparable with those for fixation times. Again, there were no significant differences in fixation counts for the areas introduction, items, and total, but there were significant differences for the rating-scale and answer field areas (Table 2). Respondents needed significantly more fixations to inspect the 5VER rating scale than both other rating scales, which is evident in Table 2. According to the results of post hoc tests, only the difference between the 5VER and 7NUM rating scales was significant ($d_{(5\text{VER/7VER})} = 10.05, p = .14$; $d_{(5\text{VER/7NUM})} = 12.55$, $p = .05$; $d_{(7\text{VER/7NUM})} = 2.04, p = 1.00$). As with fixation times, respondents needed significantly more fixations to map their responses on the answer area in the 7NUM group than in the other two groups, which is also supported by the results of the post hoc tests, $d_{(5\text{VER/7VER})} = 6.12, p = 1.00$; $d_{(5\text{VER/7NUM})}$ $= 28.45, p = .001$; $d_{(7\text{VER/7NUM})} = 22.32, p = .01$.

To summarize these results, the eye tracking study provided support for the H1 insofar as the cognitive process related to attention and effort during

Table 2. Fixation Counts on Different Areas of the Screen in the Three Rating-scale Groups.

Rating Scale	5VER		7VER		7NUM			
Screen Areas	Mean	*SD*	Mean	*SD*	Mean	*SD*	*F*(2, 79)	Part. η^2
Introduction	21.96	17.23	19.46	12.83	22.66	14.56	.32	.01
Items	162.20	66.20	162.96	100.04	156.76	62.24	.05	.00
Rating scale	46.64	15.65	36.50	21.34	34.34	38.83	3.34*	.08
Answer field	48.76	16.66	54.14	29.60	76.69	34.26	7.76**	.17
Total	291.96	92.58	285.32	143.69	302.76	91.65	.02	.00
n	25		28		29			

Note: MANCOVA results with respect to the main effects of (1) rating-scale groups: Wilks'λ= 0.69; $F_{(10,\ 148)}$= 3.09; p = .00; η^2 = 0.17; (2) respondents' baseline fixation count (covariate): Wilks'λ = 0.90; $F_{(5,\ 74)}$ = 1.06; p = .39; η^2 = 0.07. MANCOVA = multivariate analysis of covariance; 5VER = five-category verbal form; 7VER = seven-category verbal form; 7NUM = seven-category numeric form; *SD* = standard deviation.
*p < .05. **p < .01.

the mapping stage differed between the numeric and verbal rating scales. Respondents needed shorter fixation times and fewer fixation counts to comprehend the seven-category numeric rating scale than for the five-category verbal scale, whereas there was no significant difference between the seven-category verbal and numeric rating scales. Therefore, a stronger cognitive effort associated with the reading of the labels in the verbal, compared to numeric rating scales, was not observed in the case of seven categories. An important observation was that respondents needed more fixations and longer fixation times when providing their responses in the radio button field in the case of the numeric scale than for both verbal rating scales. The cognitive effort while mapping responses was therefore greater for the numeric rating scale. Because there is nothing to comprehend on the response grid (radio buttons), respondents needed longer fixation times and more fixations to find the appropriate radio button and to provide the response. It can therefore be concluded that the cognitive burden was greater when mapping the responses with numeric than with verbal rating scales.

Results of the Validation Study

Reliability assessment. Composite reliability metrics (ρ) obtained from the three experiments are presented in Table 3. Reliability metrics were taken from the CFAs, which were fitted separately in each rating-scale group and for each construct (experiment). If the goodness-of-fit of a CFA model was

Table 3. Composite Reliability (ρ) for the Three Constructs in Different Rating Sscale Groups.

	Rating Scales		
Statistics	5VER	7VER	7NUM
EU opinions			
ρ (*SE*)	.74 (.03)	.78 (.03)	.70 (.04)
95 percent CI of ρ	[.67, .83]	[.72, .83]	[.63, .78]
n	162	161	163
Parenting			
ρ (*SE*)	.75 (.04)	.77 (.03)	.69 (.05)
95 percent CI of ρ	[.69, .82]	[.70, .83]	[.60, .79]
n	163	161	164
General ideologies			
ρ (*SE*)	—	.86 (.02)	.81 (.03)
95 percent CI of ρ	—	[.82, .89]	[.75, .86]
n	—	243	245

Note. ρ = composite reliability; *SE* = standard error; CI = confidence interval; 5VER = five-category verbal form; 7VER = seven-category verbal form; 7NUM = seven-category numeric form.

not reasonable, stepwise modifications were conducted, as indicated by modification indices (MIs), starting with the highest MI. This was necessary to obtain an appropriate basis for reliability estimation in an experimental group.

For the two-factor model of the EU items, tenable goodness-of-fit (RMSEA $\leq$ .08; CFI $\geq$ .93) was obtained in each rating-scale group. However, in the 5VER group, three error covariances were included in the model to obtain this model fit; in the 7VER group, one error covariance, and in the 7NUM group, four error covariances and one cross loading were included. All items substantially and significantly loaded (standardized $\lambda \geq .27$) on the corresponding factors, except for item 4 in the 7NUM group. Therefore, as far as the quality of measurement model is concerned, it was relatively poor in the case of the numeric rating scale. The error covariances were located in the denominator of the equation, to assess reliability (Raykov 2012).

The highest reliability was obtained for the 7VER rating-scale group; the lowest for the 7NUM group (Table 3). The 95 percent confidence intervals of the reliability coefficients revealed no substantive differences between the 5VER and 7VER groups, because the confidence intervals largely overlapped. A comparison of the confidence intervals in the groups 7VER and

7NUM demonstrated that reliability tended to be higher in the 7VER group, due to only partially overlapping confidence intervals.

Now we look at the two factors for gender roles—parenting and general ideologies—used in experiments 2 and 3, respectively. In the second experiment, three groups—5VER, 7VER, and 7NUM—were compared. In the third experiment, the 7VER and 7NUM groups were included. With three items, saturated CFA models were obtained, where goodness-of-fit cannot be evaluated, but which can be used as a (perfect) base for reliability estimation when comparing factor loadings and error terms between different rating-scale groups. All items were significantly and highly loaded on the corresponding gender role factors ($\lambda > .60$).

It is evident from Table 3 that the results of experiment 2 with parenting items were comparable with those obtained with the EU items: while reliability did not differ between the rating-scale groups 5VER and 7VER, it is lower in group 7NUM than in both other groups. Due to only partly overlapping confidence intervals between the 5VER and 7NUM as well as between the 7VER and 7NUM groups, it can be concluded that reliability tended to be lower in the 7NUM group than in the other groups. This result cannot be explained by the gender shift in the experiment 2 (see the methods section), because the results are comparable with the results of the experiment with the EU items, where no such a shift was observed. In the third experiment, with the items of the gender ideologies factor, a lower reliability was obtained in the 7NUM group than in the 7VER group (Table 3).

In sum, reliability analysis obtained roughly comparable results for the three different constructs under investigation, showing that reliability tended to be lower when a numeric rating scale was used, compared to verbal rating scales. Therefore, H2 is supported by the results. In particular, the difference between the seven-category verbal and numeric rating scales can be evaluated as substantial and of practical relevance. A value of reliability of $\rho = .50$ reflects a point of reliability in which 50 percent of the entire variance is displayed by the true score variance, whereas the other 50 percent reflects the error variance. Therefore $\rho = .50$ can be considered as the lowest possible starting point for a value of reliability—an absolutely insufficient level. A reliability of $\rho < .70$ can be interpreted as inacceptable (Fisseni 1997), because less than 70 percent of the entire variance is due to the true variance. The reliability value of $\rho = .70$ can be evaluated as being low, but acceptable, whereas the value of $\rho = .78$ (as it approximates to .80) can be interpreted as a middle level. Similarly, the reliability of $\rho = .81$ (middle magnitude) differs from the $\rho = .86$, which can be interpreted as a relatively high value (close to 90 percent of the entire variance being caused by the true

Table 4. Validity Coefficients (Standardized β Coefficients of Univariate Linear Regressions) by Rating Sscale Group.

	Rating Scales		
Concepts	5VER	7VER	7NUM
Criterion: authoritarianism			
Predictors			
EU 1	.38**	.42***	.44**
EU 2	−.12	−.29***	−.10
Parenting	−.19*	−.32***	−.20**
General ideologies	—	.52***	.40***
Criterion: left–right self-placement			
Predictors			
EU 1	.16*	.26***	.20**
EU 2	−.18*	−.28***	−.10
Parenting	−.20*	−.34***	−.17*
General ideologies	—	.29***	.28***

Note. EU = European Union; 5VER = five-category verbal form; 7VER = seven-category verbal form; 7NUM = seven-category numeric form.
*p < .05. **p < .01. ***p < .001.

variance). Fisseni's (1997) rule is a very approximate rule of thumb; however, because the observed range of reliability coefficients was between $.69 \leq \rho \leq .87$, it seems to suit the given context.

Validity Assessment

Before calculating the linear regressions to evaluate validity, several prerequisites (homoscedasticity, normally distributed errors, and linearity) were tested and found to be fulfilled. A few outliers with standardized residuals greater than $z = 3.29$ were not included in the analysis. However, the results with and without outliers did not differ as far as the conclusions from the results are concerned.

To analyze concurrent criterion validity, positive associations of authoritarianism with the EU 1 factor and gender ideologies factor (conventional opinions) and negative associations with the EU 2 factor and parenting (nonconventioanl opinions) were previously stipulated (see methods section). The results obtained from the univariate linear regressions are presented in Table 4, in which the standardized linear regression coefficients are displayed.

Examination of the relationships of authoritarianism with the EU 1, general ideologies and parenting factors (see Table 4) reveals that these were expected significant relationships in each rating-scale group. However, for "general ideologies" and "parenting," the relationships were stronger for the seven-category verbal rating scale than for the other rating scales. Marked differences between the experimentally varied rating scales were observed for the associations of the EU 2 factor with authoritarianism. The expected negative relationship is significant only in the group with the verbal rating scale with seven categories (7VER) but not in the two other groups.

Concerning the right–left self-placement, high scores for the EU 1 factor (criticism of the EU) and gender ideologies (conservative opinions on gender role) were expected to be positively associated with right placement (scores were *left* = 1; *right* = 10). High scores of the EU 2 factors (approval of the EU) and of the parenting factor (rejection of the traditional gender role) were expected to be associated with left orientation. Whereas the assumption for the association between the EU1, parenting, and general ideology factors is supported in every rating-scale group (by significant regression coefficients), the assumption with respect to the EU2 factor was supported only in the two verbal rating-scale groups. The corresponding association was not significant in the case of the numeric rating scale. For parenting, the relationship is more pronounced in the seven-category verbal rating-scale group than in the other two groups.

In summary, the H3 was partly supported by the data. The conclusions concerning the validity varied according to the rating-scale design, whereas all theoretical assumptions could be supported in the 7VER group. In this group, some relationships were also more pronounced than in other groups. Particularly when the numeric rating scale was used, some hypotheses derived to test validity were not supported.

Discussion and Conclusions

This study was designed to analyze cognitive processes and measurement quality in terms of the reliability and validity of verbal and numeric rating scales, with a focus on seven-category numeric rating scales. Predominantly five to seven categories were found in previous research to be associated with the highest measurement quality Krosnick and Fabrigar (1997). While five-category numeric rating scales were found to provide insufficient measurement quality in online and non-online surveys (Menold et al. 2014; Menold and Kemper 2015; Menold and Tausch 2016), only a few studies analyzed numeric rating scales with seven categories in online mode. In addition, although eye tracking results for respondents' attention to five-category rating scales exist, no such data are available for seven-category numeric rating scales. The study continued the investigations carried out by Menold et al. (2014) and Menold and Tausch (2016) and used a construct incorporated into these studies as well as additional constructs.

With respect to the cognitive processes of respondents, the results of the present study are only partly in line with the results obtained by Menold et al. (2014), in which respondents needed considerably more attention to comprehend the five-category verbal than the five-category numeric rating scale. In the present study, respondents also needed significantly more fixations to comprehend the five-category verbal rating scale, compared to the seven-category numeric rating scale. However, this difference was not significant between the seven-category verbal and numeric rating scales. Therefore, in the case of seven categories, respondents' cognitive effort was not significantly greater for reading and comprehending verbal labels, compared to their numeric equivalents. A novel result of this study, related to the cognitive process of respondents, is that respondents needed more fixations and longer fixation times when mapping their responses onto the grid with response buttons. Therefore, even though the total time and number of fixations did not differ between the verbal and numeric rating scales, it was more difficult for the respondents to map their responses onto the response categories with the numeric than with with the verbal rating scales. Obviously, it was a burdensome task to find a response category that suited the opinion of the respondents with the seven-category numeric rating scale. This greater effort (or greater cognitive burden) required to provide a response can be explained, firstly, by the low compatibility of numeric ratings with the survey situation, due to the potential conflict between the predominant associative reasoning system and a request to provide seemingly exact responses by employing numbers (Windschitl and Wells 1996). A second explanation would be that verbal labels provide a clearer understanding of the measurement dimension and of the meaning of the intermediate categories. Satisficing behavior, in the case of numeric labels, is probably not an appropriate explanation, because respondents made a greater cognitive effort while mapping their responses and this yielded a lower measurement quality—a second important result of the study.

Combining results from the eye tracking study with the results about the measurement quality in terms of reliability and validity is a particular strength of the design used in the present research, so that cognitive process of respondents can explain the effect of rating-scale forms on reliability and validity. As far as reliability is concerned, the present experiments yielded comparable results for different constructs: reliability tended to be lower when numeric labels were used. The result obtained with respect to reliability is comparable to the results of Menold et al. (2014) and Menold and Tausch (2016) who used an agreement five-category numeric rating scale. In addition, the results are in line with the results of Krosnick and Berent (1993) who compared retest reliability between seven-category numeric and verbal rating scales using a variety of modes, samples, concepts, and measurement dimensions. A relevant contribution of the present study to research is that a cross-sectional reliability assessment method appropriate for the given data was used. Studies that used α as a cross-sectional reliability measure did not show the effects of labeling on cross-sectional reliability (Churchill and Peter 1984), but the results of those studies were probably biased, due to the use of a relatively inexact reliability metric.

Comparable to the results obtained by Windschitl and Wells (1996) and Krosnick and Berent (1993), who used different non-online data collection modes and five as well as seven categories in rating scales, criterion validity was limited for the numeric rating scale in the present study. The results have an implication for the testing of theory-driven hypotheses in sociological and social science research because social science researchers are especially interested in the prediction of variable values by other variables in regression models. The results of the present study provide clear evidence that support or reject theory-driven hypotheses about such relationships is depending upon the rating-scale format used in a study. In particular, some of the theoretically predicted relationships between opinions about the EU, on the one hand, and authoritarianism and right-left orientation, on the other, were not supported when the numeric rating scale was used, while this was not the case with the verbal rating scale with seven categories. These results demonstrate that researchers should be particularly cautious in interpreting regression model results that were obtained with numeric rating scales.

Taking the results of the present study into account, the use of fully verbalized seven categories can be recommended for online surveys, while the other—even stronger—recommendation is to avoid numeric rating scales. Because similar results were obtained for non-online modes, for other kinds of measurement continuum of rating scales, and in other countries (particularly the studies by Krosnick and Berent 1993; Windschitl and Wells 1996), as well as for five categories in online and non-online modes (Menold et al. 2014; Menold and Tausch 2016), the results presented here seem not to be restricted to the special setting of the study, such as the use of seven categories, application dimension, or a German Internet users' sample. For an online setting and a frequency rating scale, Menold and Kemper (2015) show that numeric rating scales as well as a combination of verbal and numeric labels in a rating scale led to low measurement quality. However, more insights are needed with respect to both the cognitive processes and the combination of verbal and numeric labels in a single rating scale. Additionally, to the author's best knowledge, no evidence is available about the cross-cultural comparability of verbal versus numeric rating scales. Furthermore, there are no mixed-mode or mixed device studies that focus on the cognitive process or on measurement quality in terms of reliability and validity. Since less is known about the cognitive processes of respondents, more eye tracking studies are needed.

Appendix

Response categories for all items in the Appendix are shown in Figure 1.

Items used in the eye tracking study and in experiment 1 of the validation study (Source for the translated items: Menold et al. 2014).

The following statements are about the policy of the European Union. Please indicate to what extent a statement applies.

1. In Germany, social security is weakened by European Union (EU) regulations.
2. The regions in Europe should preserve their sovereignty.
3. A member state should be able to quit the EU of its own accord.
4. The eastward expansion led to an economic upturn in Germany.
5. The eastward expansion endangered job security in Germany.
6. All EU citizens should be able to decide on EU contracts by referendum.
7. The eastward expansion led to an increase in criminal activities in Germany.
8. The introduction of the Euro has been a great success so far.
9. The Euro should be introduced into all EU states.
10. The EU needs a common foreign and security policy.

Items used in experiment 2 of the validation study (parenting items from the gender role battery; author's own English translation)

In the following statements, gender roles are addressed. Please evaluate each statement with the help of the scale provided.

1. An employed mother can have a loving and close relationship with her children, just as an unemployed mother.
2. A young child would certainly suffer, if her or his mother is working.
3. It is even better for a child if his or her mother is employed and does not only concentrate on household work.

Items used in experiment 3 of the validation study (general ideologies items from the gender role battery; author's own English translation)

In the following, more statements about gender roles are presented. Please evaluate each statement with the help of the scale provided.

1. It is more important for a woman to support her husband (partner) with his career than to build a career herself.
2. It is better for all if the male parent/father is fully involved in his professional life and the female parent/mother stays at home and takes care of the household and the children.
3. A married woman should abandon her professional life, if her career opportunities are limited, and if her husband is able to fully support the needs of the family.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

[1] http://www.gesis.org/unser-angebot/daten-analysieren/umfragedaten/european-values-study/, retrieved April 1, 2016.

[2] http://www.europeansocialsurvey.org/, retrieved April 1, 2016.

[3] http://www.gesis.org/wahlen/gles/Daten_und_Dokumente/Dokumente, retrieved April 1, 2016.

[4] http://www.gesis.org/unser-angebot/daten-erheben/gesis-panel/gesis-panel-longitudinal-core-study/, retrieved April 1, 2016.

[5] https://www.lissdata.nl/lissdata/research/liss-core-study, retrieved April 1, 2016

[6] http://www.gesis.org/allbus/allbus-home/, retrieved April 1, 2016.

References

Ackerman, Rakefet and Morris Goldsmith. 2011. "Metacognitive Regulation of Text Learning: On Screen Versus on Paper." *Journal of Experimental Psychology: Applied* 17:18-32.

Aichholzer, Julian and Eva Zeglovits. 2015. "Balancierte Kurzskala autoritärer Einstellungen (B-RWA-6)." in *Zusammenstellung sozialwissenschaftlicher Items und*

Skalen, edited by D. Danner and A. Glöckner-Rist Mannheim: GESIS. Retrieved September 7, 2017 (http://zis.gesis.org/skala/Aichholzer-Zeglovits-Balancierte-Kurzskala-autorit%C3%A4rer-Einstellungen-(B-RWA-6.)

Altemeyer, Bob. 1981. *Right-wing Authoritarianism*. Winnipeg, Canada: University of Manitoba Press.

Alwin, Duane F. and Jon A. Krosnick. 1991. "The Reliability of Survey Attitude Measurement: The Influence of Question and Respondent Attributes." *Sociological Methods & Research* 20:139-81.

Andrews, Frank M. 1984. "Construct Validity and Error Components of Survey Measures: A Structural Modeling Approach." *Public Opinion Quarterly* 48:409-42.

Beierlein, Constanze, Frank Asbrock, Mathias Kauff, and Peter Schmidt. 2014. "Die Kurzskala Autoritarismus (KSA-3): "Ein ökonomisches Messinstrument zur Erfassung dreier Subdimensionen autoritärer Einstellungen." in *Zusammenstellung sozialwissenschaftlicher Items und Skalen*, edited by D. Danner and A. Glöckner-Rist Mannheim: GESIS. Retrieved September 7, 2017 (http://zis.gesis.org/skala/Beierlein-Asbrock-Kauff-Schmidt-Kurzskala-Autoritarismus-(KSA-3).

Bendig, A. W. and J. B. Hughes, II. 1953. " Effect of Amount of Verbal Anchoring and Number of Rating-scale Categories upon Transmitted Information." *Journal of Experimental Psychology* 46:87-90.

Bernardin, H. J., Mary B. LaShells, Patricia C. Smith, and Kenneth M. Alvares. 1976. "Behavioral Expectation Scales: Effects of Developmental Procedures and Formats." *Journal of Applied Psychology* 61:75-79.

Braun, Michael. 2006. *Funktionale Äquivalenz in interkulturell vergleichenden Umfragen. Mythos und Realität*. Mannheim, Germany: ZUMA.

Breyer, Bianka. 2015. "Left-Right Self-Placement (ALLBUS)." in *Zusammenstellung sozialwissenschaftlicher Items und Skalen*, edited by D. Danner and A. Glöckner-Rist Mannheim: GESIS. Retrieved September 7, 2017 (http://zis.gesis.org/skala/Breyer-Left-Right-Self-Placement-(ALLBUS)

Churchill, Gilbert A. and J. P. Peter. 1984. "Research Design Effects on The Reliability of Rating Scales: A Meta-analysis." *Journal of Marketing Research* 21:360-75.

Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences*. New York, NY: Erlbaum.

Cronbach, Lee J. 1951. "Coefficient Alpha and The Internal Structure of Tests." *Psychometrika* 16:297-334.

Duncan, Lauren E., Bill E. Peterson, and David G. Winter. 1997. "Authoritarianism and Gender Roles: Toward A Psychological Analysis of Hegemonic Relationships." *Personality and Social Psychology Bulletin* 23:41-49.

Epstein, Seymour. 1990. "Cognitive Experiential Self-theory." Pp. 165-92 in *Handbook of personality: Theory and research*, edited by Lavrence Pervin. New York, NY: Guilford Press.

Finn, R. H. 1972. “Effects of Some Variations in Rating Scale Characteristics on the Means and Reliabilities of Ratings.” *Educational and Psychological Measurement* 32:255-65.

Fisseni, Hermann J. 1997. *Lehrbuch der Psychologischen Diagnostik*. Göttingen, Germany: Hogrefe.

Groves, Robert M., Floyd J. Fowler, Mick P. Couper, James M. Lepkowski, Eleanor Singer, and Roger Tourangeau, eds. 2004. *Survey Methodology*. New Jersey, NY: Wiley.

Krosnick , Jon A. and Duane F. Alwin. 1987. “ An Evaluation of a Cognitive Theory of Response-Order Effects in Survey Measurement .” *Public Opinion Quarterly* 51:201.

Krosnick, Jon A. and Matthew K. Berent. 1993. “Comparisons of Party Identification and Policy Preferences: The Impact of Survey Question Format.” *American Journal of Political Science* 37:941-64.

Krosnick, Jon A. and Leandre R. Fabrigar. 1997. “Designing Rating Scales for Effective Measurement in Surveys.” Pp. 141-64 in *Survey Measurement and Process Quality*, edited by L. E. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin. Hoboken, NJ: John Wiley & Sons, Inc.

Lenzner, Timo, Lars Kaczmirek, and Mirta Galesic. 2014. “Left Feels Right: A Usability Study on the Position of Answer Boxes in Web Surveys.” *Social Science Computer Review* 32:743-64.

Lord, Frederic M. and Melvin R. Novick. 1968. *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley.

McDonald, Roderick P. 1999. *Test Theory. A Unified Treatment*. Mahwah, NJ: Erlbaum.

Menold, Natalja, Lars Kaczmirek, Timo Lenzner, and Aleš Neusar. 2014. “How Do Respondents Attend to Verbal Labels in Rating Scales?” *Field Methods* 26:21-39.

Menold, Natalja and Christoph Kemper. 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* 48:431-49.

Menold, Natalja and Anja Tausch. 2016. “Measurement of Latent Variables With Different Rating Scales: Testing Reliability and Measurement Equivalence by Varying the Verbalization and Number of Categories.” *Sociological Methods & Research* 45: 678-699.

Newstead, Stephen E. and John Arnold. 1989. “The Effect of Response Format on Ratings of Teaching.” *Educational and Psychological Measurement* 49:33-43.

Parducci, Allen. 1983. “Category Ratings and the Relational Character of Judgment.” Pp. 262-82 in *Modern Issues in Perception*, edited by H.-G. Geissler, H. F. J. M. Bulfart, E. L. H. Leeuwenberg, and V Sarris. Berlin, Germany: VEB Deutscher Verlag der Wissenschaften.

Peters, David L. and Ernest J. McCormick. 1966. "Comparative Reliability of Numerically Anchored Versus Job-task Anchored Rating Scales." *Journal of Applied Psychology* 50:92-96.

Rammstedt, Beatrice, Constanze Beierlein, Elmar Brähler, Michael Eid, Johannes Hartig, Martin Kersting, Stefan Liebig, Josef Lukas, Anne-Kathrin Mayer, Natalja Menold, Jürgen Schupp, and Erich Weichselgartner. 2015. "Quality Standards for the Development, Application, and Evaluation of Measurement Instruments in Social Science Survey Research. Prepared and written by the Quality Standards Working Group." RatSWD Working Papers 245. Berlin: German Data Forum (RatSWD). Retrieved September 7, 2017 (https://www.ratswd.de/dl/RatSWD_WP_245.pdf

Rattinger, Hans, Sigrid Roßteutscher, Rüdiger Schmitt-Beck, and Bernhard Weßels. 2011. "German Longitudinal Election Study—Langfrist-Online-Tracking, T14, 23.05.-03.06.2011; Nachbefragung: 03.06.-13.06.2011." *ZA5347, Version 1.0.0*. Cologne: GESIS. Retrieved September 7, 2017 (https://dbk.gesis.org/dbksearch/sdesc2.asp?db=e&no=534).

Rattinger, Hans, Sigrid Roßteutscher, Rüdiger Schmitt-Beck, Bernhard Weßels, and Christof Wolf. 2014. "Langfrist-Online-Tracking, T17 (GLES)." *ZA5350 Datenfile Version 1.1.0*. Cologne: GESIS. Retrieved September 7, 2017 (https://dbk.gesis.org/dbksearch/SDesc2.asp?ll=10&notabs=&af=&nf=1&search=&search2=&db=E&no=5350).

Raykov, Tenko. 2001. "Bias of Coefficient α for Fixed Congeneric Measures with Correlated Errors." *Applied Psychological Measurement* 25:69-76.

Raykov, Tenko. 2012. "Scale Construction and Development Using Structural Equation Modeling." Pp. 472-92 in *Handbook of Structural Equation Modeling*, edited by Hoyle Rick H. New York, NY: The Guilford Press.

Raykov, Tenko and George A. Marcoulides. 2011. *Introduction to Psychometric Theory*. New York, NY: Taylor & Francis.

Saris, Willem E. and Irmtraud N. Gallhofer. 2007. *Design, Evaluation, and Analysis of Questionnaires for Survey Research*. Hoboken, NJ: John Wiley & Sons, Inc.

Schweizer, Karl. 2011. "On the Changing Role of Cronbach's α in the Evaluation of the Quality of a Measure." *European Journal of Psychological Assessment* 27:143-144.

Toepoel, Vera and Don A. Dillman. 2011. "Words, Numbers, and Visual Heuristics in Web Surveys: Is There a Hierarchy of Importance?" *Social Science Computer Review* 29:193-207.

Tourangeau, Roger, Mick P. Couper, and Frederick G. Conrad. 2007. "Color, Labels, and Interpretive Heuristics for Response Scales." *Public Opinion Quarterly* 71: 91-112.

Tourangeau, Roger, Lance J. Rips, and Kenneth Rasinski. 2000. *The Psychology of Survey Response*. New York, NY: Cambridge University Press.

Vygotsky, Lev S. 1934. “Thinking and Speech.” New York, NY: Plenum Press.

Weng, Li-Jen. 2004. “Impact of the Number of Response Categories and Anchor Labels on Coefficient Alpha and Test-Retest Reliability.” *Educational and Psychological Measurement* 64:956-72.

Windschitl, Paul D. and Gary L. Wells. 1996. “Measuring Psychological Uncertainty: Verbal Versus Numeric Methods.” *Journal of Experimental Psychology: Applied* 2:343-64.

Wouters, Kristel, Jeroen Maesschalck, Carel F. W. Peeters, and Marijke Roosen. 2014. “Methodological Issues in the Design of Online Surveys for Measuring Unethical Work Behavior: Recommendations on the Basis of a Split-Ballot Experiment.” *Journal of Business Ethics* 120:275-89.

Author Biography

Natalja Menold completed a Master’s degree in psychology at the University of Tuebingen in 2000. She received her doctorate from the University of Dortmund in 2006 and her habilitation from the University of Mannheim in 2017. Since 2007 she has been working at GESIS-Leibniz Institute for the Social Sciences in Mannheim, Germany. Her current position is senior project consultant and researcher, head of the team “Questionnaire Design & Evaluation”.