



# The Effect of Rating Scale Length on the Occurrence of Inappropriate Category Use for the Assessment of Job Satisfaction: an Experimental Online Study

Tanja Kutscher<sup>1,2</sup>  · Michael Eid<sup>2</sup>

Published online: 3 March 2020 / Published online: 03 March 2020

© The Author(s) 2020, corrected publication 2021

## Abstract

When job satisfaction is measured in national panel surveys using a rating scale that consists of many response categories the psychometric quality of the data obtained is often reduced. One reason lies in an inappropriate category use (e.g., in terms of response styles or ignoring superfluous categories), which occurs when respondents are faced with an overwhelmingly large number of response options. The use of response styles can also be triggered by stable respondent characteristics. The objective of the present between-subject experimental study is to explore the impact of rating scale length on the occurrence of inappropriate category use and scale reliability. In addition, this study investigates which stable respondent characteristics and job-related factors consistently predict the use of a particular response style across all experimental conditions. A sample of MTurk workers ( $N = 7042$ ) filled out a 12-item online questionnaire on aspects of job satisfaction, with a 4-, 6-, or 11-point rating scale randomly assigned. Considering the three-dimensional structure of the job satisfaction measure, we applied a multidimensional extension of the restricted mixed generalized partial credit model to explore category use patterns within each condition. The results show a similar configuration of three response-style classes in all conditions. Nevertheless, the proportion of respondents who used the rating scale inappropriately was lower in the conditions with fewer response categories. An exception was the extreme response style, which showed a similar prevalence rate in all conditions. Furthermore, we found that the use of extreme response style can be explained by a high level of general self-efficacy and perceived job autonomy, regardless of rating scale length. The findings of the study demonstrate that the prevalence of inappropriate category use can be reduced by administering rating scales with six or four response categories instead of eleven. These findings may be extended to other domains of life satisfaction.

**Keywords** Rating scale · Number of response categories · Response style · Job satisfaction · Experimental study · Mixture item response theory approach

---

**Electronic supplementary material** The online version of this article (<https://doi.org/10.1007/s41543-020-00024-2>) contains supplementary material, which is available to authorized users.

Job satisfaction (JS), as a component of subjective well-being, is a standard indicator of quality of life (Diener and Suh 1997) and is therefore one of the most studied concepts in social and organizational research. The term JS refers to an individual's contentedness with his or her job and includes subjective evaluations of relevant job aspects (e.g., income, work conditions, and relationship with colleagues) and affective states that one experiences in the workspace, such as job-related stress (Spector 1997). For organizations, high JS of employees is associated with successful human resource management, well-organized work processes, and high productivity, whereas a low level of JS indicates areas of concern that require a manager's attention (Judge et al. 2001; Tooksoon 2011). For individuals, JS is one of the important areas of life that affects individual well-being and life satisfaction (Bowling et al. 2010).

Given its high relevance, several national panel surveys measure JS at either the general or facet level. Typically, in panel surveys such as the Household, Income and Labour Dynamics in Australia Survey (HILDA; Summerfield et al. 2017), the German Socio-Economic Panel (GSOEP; Wagner et al. 2007), or the Swiss Household Panel Survey (SHP; Voorpostel et al. 2010), JS is assessed using a rating scale that consists of many response categories (e.g., an 11-point rating scale). A major problem associated with this rating scale is its susceptibility to response styles (RSs; for example, an extreme response style [ERS], a middle-category response style [MRS], acquiescence [ARS], or disacquiescence [DRS]) or other types of inappropriate category use (e.g., taking shortcuts in the form of ignoring superfluous response categories or providing careless responses; for a review, see Baumgartner and Steenkamp 2001; Van Vaerenbergh and Thomas 2013). This may be caused by an excessively large number of response categories that increase respondents' response burden and result in difficulties to discriminate between response options and to map their responses to one of the proposed response categories (Baumgartner and Steenkamp 2001; Cox 1980; Viswanathan et al. 2004). The main concern is that a high extent of inappropriate category use triggered by a long rating scale can jeopardize the validity of a JS measure, suggesting that its variance will be confounded with interindividual differences in RSs and therefore will not actually represent the interindividual variability in levels of JS (Podsakoff et al. 2003). Thus, if interindividual differences in RSs exist in the data, they will lower the data quality and can result in the questionable validity of conclusions drawn from panel studies (Baumgartner and Steenkamp 2001; De Jong et al. 2008; Morren et al. 2012). However, to date, no experimental study has compared long rating scales (e.g., an 11-point rating scale) that are frequently administrated in applied research with shorter rating scales with respect to a valid assessment of JS.

Using a between-subject experimental design, this paper explores the occurrence of inappropriate category use when assessing aspects of JS with different rating scales. In this study, we primarily consider the susceptibility of a particular rating scale to inappropriate category use as an indicator of its limited adequacy for a valid assessment of JS. This experimental design makes it possible to detect whether rating scales that consist of a few categories (short rating scales) reduce the presence of RSs, as opposed to rating scales with many response categories (long rating scales). In this study, we exclude item wording as a further source of inappropriate responses because JS measures in panel surveys typically consist of a list of aspects of the job. These items are written in plain language and do not contain any negations or extreme quantifiers (e.g., most, always or everything). In the following section, we summarize empirical evidence on the susceptibility of long rating scales to inappropriate category use. Thereafter, we present a literature review on the effects of rating scale length on RS. Due to a lack of consistent knowledge concerning another major factor affecting RS, namely a stable individual response style that is independent of the features of a

rating scale (Austin et al. 2006; Billiet and Davidov 2008; Kieruj and Moors 2013; Krosnick 1991; Van Vaerenbergh and Thomas 2013), we also address the relevance of respondent characteristics for predicting RS. Subsequently, we describe the experimental design in detail and report the results of our experimental study. We conclude with a discussion of relevant results.

## 1 The Susceptibility of Long Rating Scales to Inappropriate Category Use

First, the susceptibility of long rating scales to inappropriate category use has been demonstrated in a series of empirical studies. For example, Kutscher et al. (2017) applied mixture distribution models of item response theory (IRT) and detected at least two serious shortcomings of the 11-point rating scale used for the JS measure in the HILDA survey: (i) a large proportion of respondents adopted RSs (60%) and (ii) respondents tended to ignore many response categories because they actually expressed their attitudes using at most six out of the eleven offered categories. In contrast, using a 6-point rating scale to measure employees' satisfaction with their superiors in a similar study, Eid and Rauber (2000) found that only one-third of their respondents had a strong preference for specific response categories and tended to ignore only one category. Importantly, the majority of respondents (71%) used this rating scale in an appropriate manner. These findings illustrate that RSs may be partly avoided by shortening the rating scale. Further IRT studies have also found that short rating scales outperform their longer counterparts in terms of good coverage of the latent continuum, hierarchically ordered categories (e.g., the absence of unordered thresholds), and equidistant categories (e.g., Freund et al. 2013; Khadka et al. 2012).

Second, according to the response process model (see Tourangeau et al. 2000), respondents report their judgments by selecting the matching category of the offered response format. This represents one of the cognitive steps taken when answering items. In this step, if the administrated rating scale is not adapted to the respondents' thinking complexity and ability to discriminate, it may be a source of RSs (Baumgartner and Steenkamp 2001; Cox 1980; Viswanathan et al. 2004). According to Krosnick's concept of satisficing (1991), inappropriate responding (satisficing) of some respondents is positively related to task difficulty, which suggests that it may prove more challenging to use a rating scale with many response categories appropriately than a shorter rating scale. For example, when a rating scale is excessively long, respondents may experience increased difficulty in determining the meaning of fine categories and making decisions about which of them would adequately represent their actual judgments. This usually results in respondents' differentiation ability being overloaded and, consequently, in the use of heuristic shortcuts (e.g., focusing on a few categories and misusing labeled categories) and RSs (Greenleaf 1992a; Hamby and Levine 2016; Krosnick 1991; Swait and Adamowicz 2001; Viswanathan et al. 2004; Weathers et al. 2005). In this case, the high susceptibility of a long rating scale to inappropriate category use may be attributed to respondents' ability to appropriately differentiate among a limited number of response categories, usually up to six response options, regardless of rating scale length (Shaftel et al. 2012; Weathers et al. 2005). Individuals can differ in terms of both their differentiation ability and thinking complexity, primarily due to their cognitive abilities, experience, and educational level (De Jong et al. 2008; Miller 1956; Naemi et al. 2009; Weathers et al. 2005; Weijters et al. 2010). In particular, highly educated respondents (e.g., those found in student samples) can more accurately use rating scales with many response

categories than members of the general population (Cox 1980; Krosnick 1991). Furthermore, the use of RSs can also be related to respondents' motivation for participating in a study and their willingness to provide appropriate responses (Krosnick 1991). More specifically, highly motivated respondents may be less inclined to provide inappropriate responses when confronted with a long rating scale (Weathers et al. 2005). In contrast, when an excessively short rating scale is offered, respondents may be forced to choose a less suitable response category because a number of their subjective categories correspond to each of the broadly defined response categories (Harzing et al. 2009; Hui and Triandis 1989). To conclude, from a cognitive point of view, especially excessively long rating scales (as well as excessively short rating scales) may trigger inappropriate category use.

Third, long rating scales are often used in social research because reliability can be improved by increasing the number of response categories (Alwin and Krosnick 1991; Preston and Colman 2000; for the meta-analyses, see Churchill Jr. and Peter 1984; Saris and Gallhofer 2007). However, recent studies have stressed the risk of obtaining an artificial increase in reliability as a result of an enlarged systematic measurement error when RS effects are not eliminated from the true trait variance (Chang 1994; Jin and Wang 2014; Revilla et al. 2014; Tarka 2016). When the effects of RSs are controlled, reliability and criterion validity scores rapidly increase as a rating scale expands up to four or five response options, they remain constant for rating scales with six or seven categories, and tend to decline for rating scales with more than seven categories (Culpepper 2013; Lee and Paek 2014; Lozano et al. 2008; Maydeu-Olivares et al. 2009). This highlights that long rating scales are especially susceptible to inappropriate category use. Moreover, this effect of rating scale length on reliability is less pronounced for homogeneous scales, scales with many items, and samples consisting of highly educated respondents (Lozano et al. 2008; Maydeu-Olivares et al. 2009; Weathers et al. 2005; Weng 2004). Taken together, rating scales consisting of four to six or seven categories appear to have better psychometric properties compared to excessively short rating scales (e.g., a 3-point scale) and long rating scales (e.g., with eight or more response categories).

## 2 Effect of Rating Scale Length on the Use of Response Styles

Empirical evidence suggests that features of rating scales affect the extent of RSs that will be present in the data obtained (Cabooter et al. 2016; Hamby and Levine 2016; Kieruj and Moors 2010; Moors 2008; Moors et al. 2014; O'Muircheartaigh et al. 1999; Tourangeau et al. 2007; Weijters et al. 2010). Regarding rating scale length, Harzing et al. (2009) compared differences in common RSs when respondents were asked to answer questions concerning job values using two short rating scales (consisting of five and seven response categories, respectively). By calculating a sum-score index for a specific RS, they demonstrated that increasing the number of response categories attenuated the ERS and MRS but increased the ARS and DRS. Consistent findings have been reported in further experimental studies, the majority of which focused on the ERS by measuring diverse scales of personality traits, beliefs, and subjective well-being and manipulating rating scale length, generally between four and (seven or) ten categories (e.g., Hui and Triandis 1989; Weijters et al. 2010). In two elaborate experimental studies conducted by Clarke (2000a, 2000b), the extent of the ERS was calculated on the basis of a set of low intercorrelated items across a wide range of rating scales (including 3–10 response options). He found that, when the number of response categories was increased, short and long rating scales were affected differently: for short rating scales (3–5 options), an

increase produced an overall tendency toward a strong decrease in ERS use; for long rating scales (5–10 options), only a slight reduction was reported. Taken together, these findings illustrate that excessively short rating scales (< four response options) trigger a high extent of RSs, suggesting that respondents encounter the problem of mapping their judgments onto one of the broadly defined response categories. Instead, it may be appropriate to increase the number of response categories in short rating scales (up to seven options). This could allow for a rating scale to be refined to maximize its potential in terms of information transmission (Cox 1980) and improve its psychometric quality.

However, rating scales should not be excessively long. This is in line with the finding of an experimental study conducted by Kieruj and Moors (2010), in which both short and long rating scale lengths were manipulated (5- to 7-point and 9- to 11-point rating scales). Applications of a latent class confirmatory factor model with three content factors and one RS factor indicated that the MRS was not observed in the conditions that considered short rating scales but emerged when long rating scales were offered. Therefore, when confronted with an even-point rating scale, the respondents selected a category that was nearest to the middle of the scale as an alternative middle option. In fact, endorsing the middle category does not commonly reflect a moderate trait level. When the middle category is included in a rating scale, it has a high potential to be misused, primarily by respondents who refuse to answer, provide ambivalent or unsure responses, or do not understand the item content (Kulas and Stachowski 2009, 2013). In contrast, for a small minority of respondents with moderate trait levels, the inclusion of a middle category may prove beneficial, as they will not be forced to choose one of the adjacent categories (Hernández et al. 2004; Presser and Schuman 1980; Sturgis et al. 2014). A further argument for the inclusion of a middle category may be the higher reliability of an odd-numbered rating scale when compared with that of an even-numbered one (e.g., Borgers et al. 2004; O’Muircheartaigh et al. 1999). In summary, long rating scales make it more difficult for respondents to accurately map their responses to one of the offered response categories and therefore increase the risk of RS use as an adjustment strategy.

### 3 Respondent Characteristics for Predicting the Use of Response Styles

Although little is known about RS use that consistently occurs across different scales and over time, it may be considered a type of substantial personality disposition. The majority of research relating RS use to interindividual differences in personality traits, cognitive ability, and socio-demographic variables has yielded mixed and inconclusive findings. For example, ERS use, which has been widely examined due to its permanent occurrence in data, was not found to exhibit any consistent personality profile. In particular, some studies have found that respondents who prefer extreme categories are high in extraversion (Austin et al. 2006; Gerber-Braun 2010; Kieruj and Moors 2013; Meiser and Machunsky 2008), low in conscientiousness (Zettler et al. 2016), high in neuroticism (Baumgartner and Steenkamp 2001; Hernández et al. 2004), and low in openness to experience (Meiser and Machunsky 2008). Other studies have obtained contrasting results and reported that the ERS is positively related to conscientiousness (Austin et al. 2006; Gerber-Braun 2010), negatively linked to neuroticism (Gerber-Braun 2010), and not associated with openness to experience and agreeableness (Austin et al. 2006; Meiser and Machunsky 2008). Moreover, for ERS use, research has highlighted the relevance of intolerance of ambiguity and simplistic thinking, which were

found to account for a nearly 25% of the variance in extreme responses after controlling for gender, ethnic minority status, and level of cognitive ability (Naemi et al. 2009). Furthermore, inconsistent findings exist concerning the relationship between the ERS and cognitive abilities of respondents (for positive effect, see Gerber-Braun 2010; for no effect, see Naemi et al. 2009). Although the effect of socio-demographic variables on ERS use has been the focus of the majority of previous research, the same inconsistent picture emerged. Some empirical evidence indicates that the presence of the ERS is high for low-educated respondents (Eid and Rauber 2000; Weijters et al. 2010b), whereas other studies have found education to have no effect (Kieruj and Moors 2013; Moors 2008). With regard to gender, females have been found to have higher levels of the ERS than males in some studies (Austin et al. 2006; De Jong et al. 2008; Weijters et al. 2010b), while no gender differences were found in other studies (Clarke 2000a, 2000b; Greenleaf 1992b; Kieruj and Moors 2013; Moors 2008; Naemi et al. 2009). Some empirical evidence suggests that age is curvilinearly related to the ERS, indicating that young and older respondents are more inclined to the ERS (Baumgartner and Steenkamp 2001; De Jong et al. 2008), whereas other evidence indicates that older people tend to use the ERS more frequently than young people (Greenleaf 1992b; Kieruj and Moors 2013; Moors et al. 2014; Weijters et al. 2010b). It has also been found that only young people prefer the ERS (Austin et al. 2006; Gerber-Braun 2010). In addition, age may have no effect on ERS use (Moors 2008). Similarly, other traditional RSs have been rarely examined as compared to the ERS, thus creating an inconsistent picture concerning individuals' use of RSs. In summary, these inconsistent findings emphasize the need for a systematic study of the variables characterizing respondents who tend to use a particular response style regardless of the rating scale.

#### 4 Research Questions and Expectations

The present research examines the effects of rating scale length on category use for the assessment of JS, an important indicator of quality of life, in a between-subject experiment that varied the number of response categories: a long rating scale with 11 categories (corresponding to national panel surveys) and two short rating scales with four or six options. These short rating scales were selected in accordance with the recommendations of previous research. A rating scale with four response categories is held to be adequately short, whereas a rating scale with six response options is considered to be reasonably long. These rating scales also exclude the middle category, which is often misused (primarily by respondents who refuse to answer correctly). Thus, we assumed that the experimental conditions would differ in terms of task difficulty in such a manner that respondents in the 11-category condition would find it more difficult to answer JS items appropriately. Within each experimental condition, we also focused on the psychometric quality of the JS scale in the form of reliability. In contrast to previous research, which has mostly examined the effect of rating scale length on a few a priori-defined RSs, we explored the category use patterns present in the data and identified the number of response categories that respondents actually used while answering JS items under experimental conditions. In particular, the mixture polytomous item response theory (IRT) model was expected to provide insight into how RSs change as a result of varying rating scale length. Given previous research findings concerning this issue reported above, we formulated certain assumptions regarding the various effects of rating scale length on category use. First, compared to a long rating scale, short rating scales were expected to lead to a reduced presence

of inappropriate category use. Second, we expected that the ERS would be present in our data regardless of rating scale length. Hence, this study was expected to provide insight into how the structure of category use patterns can change depending on the proposed rating scale length.

Due to the inconsistency of previous findings, another important goal of this study was to systematically examine and identify stable respondent characteristics (e.g., personality traits, cognitive ability, and socio-demographic factors) and contextual factors (e.g., job-related factors) that can consistently predict the RSs found at different rating scale lengths. Therefore, in contrast to previous research, which has mostly focused on a few selected predictors, this study accounts for predictors previously found to be relevant for explaining RSs. The results of this study represent a valuable contribution to researchers and practitioners intending to collect data that are characterized by a negligible amount of method-related RS bias and high reliability.

## 5 Methods

### 5.1 Sample and Procedure

Data collection for the split-ballot experiment was conducted on Amazon's Mechanical Turk (MTurk) platform in the period between February and July 2015. The MTurk is an online crowdsourcing labor market where online respondents (MTurk workers) complete various tasks, so-called human intelligent tasks (HITs), for rather low pay. This research platform provides a more diverse population of respondents than student samples or otherwise recruited online samples and creates facilities for obtaining high-quality data rapidly, anonymously, and cost-effectively (Buhrmester et al. 2011; Mason and Suri 2012). In this study, MTurk workers were instructed to fill out an online questionnaire should they have provided their consent to participate in the study and met the following inclusion criteria: being at least 18 years old, being employed, and living in the USA. Additionally, to reduce the risk of satisficing response behavior, we followed the recommendations by Peer et al. (2014) and predefined within the MTurk platform that only experienced workers (at least 100 approved HITs) with a high approval rate (at least 95%) were eligible to participate in the study. The online questionnaire included JS items, personality measures, cognitive tasks, and job-related and socio-demographic questions. It was created using the software package SoSci Survey, a free tool for conducting online surveys. To avoid multiple participations, we also integrated a filter based on the MTurk IDs of workers who had already completed the study. The average response time was 16.35 min ( $SD = 5.31$ ;  $Md = 15.67$ ,  $Q_1 = 12.72$ ,  $Q_3 = 19.27$ ). The study participants received US\$ 0.50 as remuneration for participation.

The entire sample consisted of 7042 MTurk workers. Forty-three individuals did not provide any valid values on the items of job satisfaction and were excluded from the analysis sample. According to current simulation findings (e.g., Cho 2013; Huang 2016; Jin and Wang 2014; Kutscher et al. 2019), this sample size (namely 2000–2500 individuals per condition) should be sufficient for the application of the mixed polytomous IRT model to display optimal performance. Women comprised more than half of the entire sample (61%). The mean age of the respondents was 34.01 years ( $SD = 11.09$ ,  $max = 82.00$ ). The sample included a few non-native English speakers (3%). With regard to education, 9% of the respondents reported having achieved the lowest education level (the majority of them had finished high school), 27% had

completed a college degree, 48% had a graduate degree, and 16% held a postgraduate degree. The majority of respondents were employees (90%), 9% were self-employed, and 1% were unpaid or voluntary workers. Full-time employment was the dominant employment status (73%).

## 5.2 Experimental Design

To examine the impact of rating scale length on the participants' category use, we implemented a randomized between-subject design with three experimental conditions. In each of these conditions, the JS items were presented with a different number of response categories. We contrasted a long rating scale (a) with 11 categories with two shorter rating scales including (b) 6 and (c) 4 categories, respectively. Following the approach adopted from national panel surveys, we used endpoint-labeled rating scales for the JS items in all conditions. The lowest category was labeled as "totally dissatisfied", and the highest category was labeled "totally satisfied". Numerical values were presented in ascending order from the left end to the right end of the response format starting at zero (e.g., for the 11-point rating scale, from 0 to 10).

Due to randomization, respondent characteristics that may lead to inappropriate responses (e.g., respondents' motivation, cognitive ability, or specific personality traits related to RSs) were expected to be equally distributed among the experimental conditions. We examined the effectivity of randomization by comparing the experimental groups using socio-demographic and relevant job-related variables (Table 1). The experimental groups did not differ in terms of age ( $F(2, 6995) = 0.05, p = .951$ ), gender ( $\chi^2(2) = 4.35, p = .114$ ), education ( $\chi^2(6) = 4.90, p = .557$ ), employment type ( $\chi^2(4) = 4.87, p = .301$ ), or leadership level ( $\chi^2(8) = 6.95, p = .542$ ). These results were obtained by means of a univariate analysis of variance

**Table 1** Comparison of sample characteristics among the experimental conditions

	11-category condition ( <i>n</i> = 2322)	6-category condition ( <i>n</i> = 2364)	4-category condition ( <i>n</i> = 2313)
Age	34.08 (10.95)	34.05 (11.13)	33.98 (11.19)
Gender (female)	62	62	59
Marital status (married)	97	97	97
Non-native English speaker	03	03	03
Educational level			
- high school or less	09	09	09
- college	27	26	28
- graduate	48	50	47
- postgraduate	16	16	16
Employment type			
- own business	08	08	10
- employee	87	88	86
- others (e.g., voluntary work)	05	04	05
Job position			
- manager	17	15	16
- professional	28	29	27
- technician, marketing, personal service worker	15	16	16
- administrative worker	16	17	17
- service worker, Turk worker	24	23	24
Part-time occupation	26	29	28

*Note.* Mean (standard deviation) or percentage are reported



(ANOVA) for the continuous variable age and a Pearson's chi-square test for the nominal variables.

### 5.3 Measures

The questionnaire consisted of two parts: (1) the JS items, which according to the experimental design, were measured with rating scales of various lengths, and (2) four sets of measures of personality traits, cognitive tasks, and job-related and socio-demographic variables. These four sets of measures were included in the questionnaire to explain the use of the detected RSs and were identically applied in all experimental conditions. All measures are described below.

**Job satisfaction.** Respondents were asked to evaluate their level of satisfaction with various aspects of their current jobs by responding to 12 items (e.g., "Your total pay."; "The hours you work."; and "The work atmosphere and relations with your co-workers."). With the goal of covering a broad spectrum of job aspects, we intentionally did not use a standardized JS scale but instead adopted individual JS items from national panel surveys (e.g., the SHP and the HILDA surveys). Respondents answered the JS items with an 4-, 6-, or 11-point rating scale depending on the experimental condition to which they were randomly assigned.

**Big Five.** The short version of the Big Five Inventory (BFI-10; Rammstedt and John 2007) was used to measure the five personality dimensions: extraversion, neuroticism, openness to experience, conscientiousness, and agreeableness. This measure exhibits acceptable psychometric properties (Rammstedt and John 2007). The dimensions consist of two prototypical items in the form of short phrases or adjectives (e.g., "gets nervous easily" for neuroticism or "is generally trusting" for agreeableness). Respondents were asked to rate how well the statements described their personality on a 5-point rating scale from 1 (*disagree strongly*) to 5 (*agree strongly*). In each dimension, one of the items was negatively formulated and was recorded before calculating dimension scores. In the present study, the five-dimensional structure of this personality questionnaire showed an appropriate approximate fit when analyzed by confirmatory factor analysis (CFA;  $\chi^2(25) = 864.76$ ,  $p < .001$ ; RMSEA = .07, 90%-CI [.07; .07]; CFI = .91; SRMR = .05; see Table S1 in the supplementary material for details). The reliabilities of the subscales were acceptable with regard to short subscale length (McDonald's  $\omega = .68$ , .66, .50, .56, and .52 for extraversion, neuroticism, openness to experience, conscientiousness, and agreeableness subscales, respectively).

**The General Self-Efficacy Scale (GSE; Schwarzer and Jerusalem 1995)** is a widely used unidimensional self-report scale for measuring one's confidence in his or her ability to cope with demanding, stressful, or novel situations (e.g., "I can solve most problems if I invest the necessary effort." or "If I am in trouble, I can usually think of a solution."). Respondents reported to the 10 statements using a 4-point rating scale with the following labels: 1 (*not at all true*), 2 (*hardly true*), 3 (*moderately true*), and 4 (*exactly true*). In this study, the unidimensionality of the GSE scale was validated using a one-factor CFA model, which indicated a good model fit ( $\chi^2(35) = 279.04$ ,  $p < .001$ ; RMSEA = .03, 90%-CI [.03; .04]; CFI = .99; SRMR = .03; see Table S2 in the supplementary material for details). The reliability (McDonald's  $\omega$ ) was .89.

**Tolerance for Ambiguity (TA)** was assessed with a set of six items selected from the original Ambiguity Tolerance Scale (AT-20; MacDonald Jr. 1970), using the magnitude of their item-total correlations above .50 as a cutoff. The purpose was to reduce the time required to respond to the questionnaire. All items were negatively formulated and measured a general tendency to perceive or interpret ambiguous information and unstructured situations as

desirable (e.g., “I don’t like to work on a problem unless there is a possibility of coming out with a clear-cut and unambiguous answer.”). After recoding, a high score represented a high level of ambiguity tolerance. The 5-point rating scale ranged from 1 (*not at all true*) to 5 (*exactly true*). Similarly to the original scale, the short TA scale showed a unidimensional latent structure: the one-factor CFA model demonstrated an acceptable approximate model fit ( $\chi^2(9) = 366.97$ ,  $p < .001$ ; RMSEA = .08, 90%-CI [.07; .08]; CFI = .94; SRMR = .05). See Table S3 in the supplementary material for details. The reliability was acceptable (McDonald’s  $\omega = .73$ ).

**Need for Cognition.** A set of 10 items was used to measure an individual’s tendency to engage in and enjoy cognitively demanding activity. These items were selected from the original 18-item need for cognition scale (NCS; Cacioppo et al. 1996) based on whether they had their item-total correlations of at least .60 (see Cacioppo and Petty 1982). The unidimensional underlying structure of the 10-item version was validated in this study using CFA ( $\chi^2(35) = 862.07$ ,  $p < .001$ ; RMSEA = .06, 90%-CI [.06; .06]; CFI = .98; SRMR = .05; see Table S4 in the supplementary material for details). Respondents were asked to rate how well the statements, such as “The idea of relying on thought to make my way to the top appeals to me.” or “I like tasks that require little thought once I’ve learned them.” (recoded), applied to them. The 5-point rating scale was labeled as follows: 1 (*extremely uncharacteristic*), 2 (*somewhat uncharacteristic*), 3 (*uncertain*), 4 (*somewhat characteristic*), and 5 (*extremely characteristic*). McDonald’s  $\omega$  for the reduced scale was .89.

**The Decisiveness scale** (Naemi et al. 2009) is a unidimensional eight-item measure to assess a dispositional tendency to make decisions quickly, as opposed to postponing decision-making due to fear of making errors (e.g., “When faced with a problem I usually see the one best solution very quickly.”). Respondents evaluated their level of decisiveness using a 5-point rating scale ranging from 1 (*not at all true*) to 5 (*exactly true*). In the present study, the unidimensionality of this scale could be verified by carrying out a one-factor CFA model ( $\chi^2(20) = 429.55$ ,  $p < .001$ ; RMSEA = .05, 90%-CI [.05; .06]; CFI = .98; SRMR = .04; see Table S5 in the supplementary material for details). The reliability score (McDonald’s  $\omega$ ) was .83.

**Social Desirability** was assessed using a short version of the Balanced Inventory of Desirable Responding (Winkler et al. 2006), which measures two dimensions of social desirability, self-deceptive enhancement (SDE) and impression management (IM), using three items per dimension. The self-deceptive enhancement subscale includes items concerning an unconscious tendency to distort one’s perception of reality in an optimistic manner to protect one’s self-concept and self-esteem (e.g., “My first impression of people usually turns out to be right.”). The impression management subscale refers to one’s tendency to deliberately mislead other people in order to provide them with a most favorable impression of the respondent (e.g., “There have been occasions when I have taken advantage of someone.” - recoded). Respondents rated items using a 7-point rating scale ranging from 1 (*not true*), via 4 (*somewhat*) to 7 (*very true*). Before calculating the subscale scores, negatively formulated items were recoded so that higher values represented a high level of self-deceptive enhancement or impression management. The two-factor CFA model performed in this study showed a mediocre model fit ( $\chi^2(8) = 674.53$ ,  $p < .001$ ; RMSEA = .11, 90%-CI [.10; .12]; CFI = .90; SRMR = .06; see Table S6 in the supplementary material for details). The reliabilities of the two subscales were acceptable when taking into consideration the short length of these subscales (McDonald’s  $\omega = .80$  and .61 for the SDE and the IM subscales, respectively).

**Verbal memory ability** was assessed using the 10-question Verbal Memory Measure, which is a part of the Intelligence Structural Test (the English version of IST 2000 R; Beaucecel 2010). At first, respondents were requested to memorize five sets of words within one minute. Each set of words included a generic term and up to three subordinate terms (for example, “SPORT: Golf, Motorsports”). Thereafter, the respondents were posed questions (e.g., “The word with the initial letter – M – was a/an ...?”) that prompted them to select only one of the generic terms (sports, food, city, profession, and building) that referred to a subordinate term (e.g., Motorsports) asked in a particular question. The number of correct answers provided by each respondent was counted. According to the applied Rasch model, all items of the short scale showed an acceptable item fit (the weighted mean square [WMNSQ; Wright and Masters 1982] varied between 0.89 and 1.14). The one-dimensionality of this scale was supported by the small average absolute residual correlations for the items ( $aQ_3$  values [Yen 1993] ranged from 0.03 to 0.08; for details, see Table S7 in the supplementary material). The Kuder-Richardson reliability coefficient was .81.

**Verbal analogy task.** To assess the respondents’ levels of verbal competence, we used 10 questions randomly selected from the 20-question Verbal Analogy Test included in the Intelligence Structural Test (the English version of IST 2000 R; Beaucecel 2010). Each item included three terms, and specific relationships existed between the first two terms. Respondents were asked to identify this relationship and select from five given alternatives one that most strongly represented a similar relation to the third term (e.g., forest: trees = meadow:? Response alternatives: grass, hay, feed, green, pasture. Correct answer: grass). A high proportion of correct answers represented a high level of verbal competence. Applying the Rasch model, the items demonstrated a satisfactory item fit (WMNSQ values were between 0.91 and 1.08). Due to the quite small average absolute residual correlations for the items ( $aQ_3$  values varied between 0.02 and 0.05), the one-dimensionality of this scale could be supported (see Table S8 in the supplementary material for details). The Kuder-Richardson reliability coefficient was .51.

**Relevance of job.** To measure the relevance of having a job, respondents were asked to rank six life domains (health, finances, job, family, free time and friends, and home and living environment) in order of importance on a scale of 1 (*the most relevant*) to 6 (*the least relevant*). The position of having a job in the hierarchy indicated its degree of relevance.

**Job characteristics.** The 10-item measure was adopted from the first wave of the HILDA survey (Summerfield et al. 2017) to assess the respondents’ perception of four aspects of psychosocial work conditions: job autonomy (e.g., “I have a lot of freedom to decide how I do my own work.”), job skills (“My job often requires me to learn new skills.”), job-related stress (“I fear that the amount of stress in my job will make me physically ill.”), and job security (“I have a secure future in my job.”). Respondents were asked to rate items using a 7-point rating scale from 1 (*strongly disagree*) to 7 (*strongly agree*). The four-dimensional underlying structure of the JC measure, which was previously described by Kutscher et al. (2017), was verified in this study by carrying out CFA, which indicated an acceptable approximate model fit ( $\chi^2(29) = 1665.85$ ,  $p < .001$ ; RMSEA = .09, 90%-CI [.09; .09]; CFI = .93; TLI = 0.90; SRMR = .07; for details, see Table S9 in the supplementary material). The internal consistency of the subscales was acceptable (McDonald’s  $\omega = .82, .72, .80$ , and  $.65$  for job autonomy, job skills, job-related stress, and job security subscales, respectively).

**Further job-related variables.** Respondents were asked to report their *employment type* (1 = “own business”, 2 = “employee”, 3 = “others, e.g., voluntary work”), *job position* (1 = “level 1: manager, self-employed, etc.”; 2 = “level 2: professional, researcher, etc.”; 3 = “level

3: technician, marketing, personal service worker, etc.”; 4 = “level 4: administrative worker, etc.”; 5 = “level 5: service worker, machinery operator, MTurk worker, etc.”), *tenure at current position* (1 = “less than 1 year”, 2 = “1–3 years”, 3 = “4–6 years”, 4 = “7–9 years”, 5 = “10 years or longer”), *full- or part-time occupation*, and the *size of organization* in which they work (1 = “small: less than 50 employees”, 2 = “middle: 50–200 employees”, 3 = “large: 200 employees or more”).

**Socio-demographic variables.** Finally, respondents reported their *age* (in years), *gender* (1 = “female”, 2 = “male”), *educational level* (1 = “high school graduate or less”, 2 = “some college,” 3 = “associate degree or bachelor’s degree”, 4 = “master’s degree, doctorate or professional degree”), *married status* (1 = “married”, 2 = “unmarried”), and *first language* (1 = “English”, 2 = “other language”).

## 5.4 Statistical Analyses

**Multidimensional restricted mixed IRT model.** To analyze individual differences in RSs with respect to the JS items, the mixture IRT approach was applied within each of the experimental conditions (for an overview, see von Davier and Yamamoto 2007). This approach assumes the existence of latent classes of respondents with homogeneous response patterns. But within latent classes, respondents are allowed to vary in their latent trait values. Therefore, it is possible that individuals in different classes have the same value on the latent trait variable but differ in their response styles. Because a factor analysis revealed that the JS items considered in this study had a three-factor structure (see the Results section), we used a multidimensional extension of the restricted mixture generalized partial credit model (rmGPCM) as a target model in all experimental conditions (for more details on the rmGPCM, see Kutscher et al. 2017). In the multidimensional model, an rmGPCM is assumed for each of the three JS subscales consisting of unique sets of items, and the three latent trait variables are allowed to correlate. Furthermore, in the model, based on previous evidence suggesting that peoples’ use of response categories is relatively stable across traits or attitudes in large-scale assessment studies (e.g., Weijters et al. 2010a; Wetzel et al. 2013; Zettler et al. 2016), it was assumed that the number of latent classes with a specific category use pattern is the same across latent dimensions. Thus, the model allows identifying latent classes with RSs that are common across all JS subscales. Within a latent class, the multidimensional rmGPCM defines the response probability of an item category across the entire latent continuum as a logistic function of two types of item parameters: class-specific threshold parameters (denoting transition points between two adjacent categories) and a discrimination parameter. In the multidimensional rmGPCM, it is assumed that the discrimination parameter of an item is invariant across latent classes. This assumption is reasonable in the context of exploring response patterns and makes it possible to reduce the complexity of the model and to prevent the occurrence of estimation problems.

The multidimensional rmGPCM is defined by the following equation:

$$P_{\text{vtix}}(\theta) = \sum_{g=1}^G \pi_g \frac{\exp \left[ \sum_{s=0}^x \delta_{it} (\theta_{\text{vtg}} - \tau_{\text{istg}}) \right]}{\sum_{c=0}^m \exp \left[ \sum_{s=0}^c \delta_{it} (\theta_{\text{vtg}} - \tau_{\text{istg}}) \right]} \quad (1)$$

where  $P_{\text{vtix}}(\theta)$  denotes the probability of obtaining a response in a category  $x$  ( $x \in \{0, \dots, m\}$ ) to a

categorical item  $i$  (belonging to a dimension  $t$ ) for a respondent  $v$  assigned to a latent class  $g$  with a latent trait value  $\theta_{vtg}$  on a continuous latent dimension  $t$ . It is assumed that the latent trait variables are normally distributed with a mean of zero and the latent variances are freely estimated within each latent class  $g$ .  $\delta_{it}$  is a class-invariant discrimination parameter of item  $i$  belonging to dimension  $t$  (with  $\delta_{it} > 0$  and with  $\delta_{it} = 1$ ), and  $\tau_{istg}$  is a class-specific threshold parameter of item  $i$  with respect to dimension  $t$  (with  $s \in \{0, \dots, c\}$  and  $\tau_{i0tg} = 0$  for all  $i$  of all  $t$  in all  $g$ ). The model does not directly estimate the number of latent classes, but they can be determined by comparing models with a different number of latent classes using goodness-of-fit statistics. In addition, the model allows to estimate the sizes of latent classes ( $\pi_g$ , with  $\sum_{g=1}^G \pi_g = 1$ ). The marginal maximum likelihood estimation of the model parameters was obtained using the computer program Latent GOLD 5.0 (Vermunt and Magidson 2013). (For the model script and estimation settings, see part B of the supplementary material.)

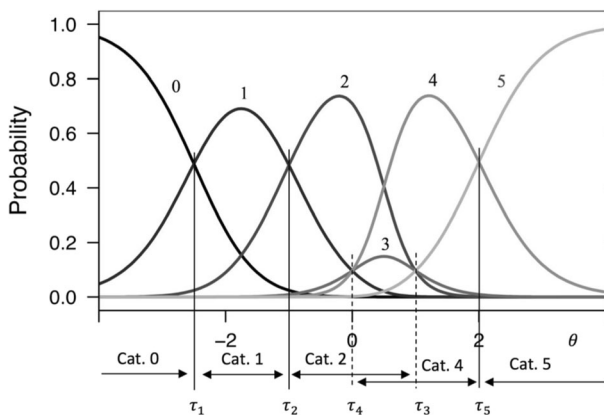
Within each experimental condition, we estimated the multidimensional rmGPCM including up to five latent classes and determined the best-fitting solution using the Bayesian information criterion (BIC; Schwarz 1978), which works well and is consistent in the context of complex models and large sample sizes (Dziak et al. 2012). We purposely chose neither the Akaike's information criterion with the tripled number of model parameters (AIC3; Bozdogan 1994) nor the sample-size adjusted BIC (SABIC; Sclove 1987), both of which showed a good performance for model selection in unidimensional polytomous IRT models (Kutscher et al. 2019). However, there is a lack of evidence concerning their performance for multidimensional IRT models. The lowest BIC value indicates the model with the adequate number of latent classes. We could not evaluate the absolute model fit of the estimated models using test statistics to compare the expected and observed frequencies of response patterns (e.g., Pearson and Cressie-Read  $\chi^2$  goodness-of-fit statistics or the likelihood-ratio test) because of the disproportionately high number of potential response patterns (e.g., in the 11-category condition, 3,138,428,376,721 possible response patterns for the twelve items) in relation to the sample size used in each of the experimental conditions (range: 2313–2364 individuals). In addition, bootstrapping the distribution of the Pearson and Cressie-Read  $\chi^2$  goodness-of-fit statistics was unrealistic due to excessively lengthy computing times caused by the complexity of the model (Nylund et al. 2007). Furthermore, alternative goodness-of-fit assessment methods (as described by Maydeu-Olivares 2013) were considered less promising due to a large number of response categories, the model complexity, and the sparse table problem (Maydeu-Olivares and Joe 2008). Therefore, we reported the sum of bivariate residuals for each class solution. A low value indicates a high degree of similarity between the estimated and observed bivariate associations of the variables (Vermunt and Magidson 2013).

**Exploring category use patterns.** To interpret category use patterns in latent classes within an experimental condition, we plotted the class-specific response probabilities of the item categories in the form of category characteristic curves (CCCs) using item parameters from the best-fitting model. As depicted for a fictitious item with six response categories ( $x = 0, \dots, 5$ ) in Fig. 1, the response probability for the first and the last categories monotonically decreases and increases, respectively. Thus, it is very likely that the first category would be endorsed at the lower area of the latent continuum. The opposite holds true for the last category. The CCCs of the other categories are unimodal, with their peaks highlighting the corresponding segments of the latent continuum at which a certain category has the highest probability of being endorsed. The threshold parameters are located on the latent continuum according to their increasing difficulty and represent the intersection points of the CCCs of two categories,  $x - 1$  and  $x$ . First,

we examined the order of threshold parameters as a relevant indicator of item functioning. In an ideal case, the threshold parameters would be ordered ( $\tau_{i,s-1} < \tau_{i,s}$ ), and each response category therefore has an area on the latent variable in which it is more strongly preferred (it has a higher response probability than the other categories). If two thresholds are unordered (Fig. 1), the response probability of the concerned category will always be lower than the response probabilities of all other categories, and, as a consequence, this category will be avoided (Andrich 2010; Smith Jr et al. 2011; Wetzel and Carstensen 2014). For the fictitious item, the 3rd and 4th thresholds are unordered ( $\tau_{i3} > \tau_{i4}$ ), indicating that the respondents tended to ignore category  $x = 3$ . Second, we determined the respondents' preference for categories by evaluating the distances between adjacent thresholds. These distances represent the widths of corresponding categories on the latent continuum, with large category width being associated with a more preferred category. Thus, the large widths of extreme categories combined with the small widths of middle categories correspond to the ERS; the opposite holds for the MRS. Figure 1 represents the case of an item with nearly equidistant categories.

Generally, the magnitude of the item discrimination parameter affects the widths of the categories and their response probabilities. For an item with a high discrimination parameter, the CCCs are steeper and narrower. Consequently, the categories are somewhat smaller but possess a higher probability of being endorsed compared to an item with a lower discrimination parameter.

**Detecting careless responses.** Following recommendations by Curran (2016) intended to ensure that researchers obtain high-quality online data, we used four screening techniques to identify respondents exhibiting different types of careless responses: (i) an attention check designed to detect inattentive respondents who failed to read items carefully over the course of the survey; (ii) response time, indicating whether respondents spent the minimum amount of time required to answer accurately (e.g., to determine if respondents engaged in quick responding); (iii) a long-string index assessing a respondent's tendency to select the same response options for many consecutive items (invariant responders); and (iv) a resampled intraindividual reliability (RIR) score that indicates whether a respondent provided consistent responses within several measures and makes it possible to detect random responding. For the attention check, four additional items that explicitly instructed respondents to indicate a particular type of response were incorporated at various points throughout the survey (e.g.,



**Fig. 1** Category characteristic curves (CCCs) for a fictitious item with six response categories and an unordered threshold (threshold parameters:  $\tau_1 = -2.5$ ,  $\tau_2 = -1$ ,  $\tau_3 = 1$ ,  $\tau_4 = 0$ ,  $\tau_5 = 2$ ; discrimination parameter  $\delta = 2$ )

“In order to verify that the program retains the data correctly, please select the option ‘strongly agree’ for this statement.”). An incorrect response to at least one of these items indicated a respondent’s failure to devote sufficient attention to his or her responses. Respondents’ response time was recorded using the built-in timer of the SoSci Survey tool. We considered respondents as investing insufficient effort when their response times were faster than the cutoff value, which equaled the mean value minus two standard deviations of the logarithmized response time variable. The presence of long strings in each respondent’s responses was determined using predictor scales. Because the long-string index is a scale-dependent statistic, we defined respondents’ responses as careless when they included invariant responses to more than 75% of the items on a scale for more than two scales. An individual RIR score was calculated as an average within-person correlation between two vectors containing his or her mean values for two sets of items that were randomly selected from one of all four unidimensional predictor scales repeatedly. In this calculation, we used z-transformed item scores to overcome differences in rating scales. Low response consistency in a respondent’s values (e.g., an RIR score below 0.3) indicated random responding. In addition, the results of this screening should have indicated what types of careless responses were present in latent classes with different RSs.

**Predicting latent class assignment.** Within each experimental condition, respondents were assigned to the latent classes for which their assignment probability was maximum. Multinomial logistic regression was applied to predict the posterior assignment to latent classes from the best model solution by means of socio-demographic variables, personality traits, cognitive ability to process information, and job-related variables. For categorical predictors (e.g., job position), sets of dummy variables were included in the analysis. We used the adjusted three-step method implemented in Latent GOLD 5.0 to remove the impact of a classification error that resulted from applying the multidimensional rmGPCM to regression coefficients and standard errors (Vermunt and Magidson 2013).

## 6 Results

### 6.1 Descriptive Analysis for Job Satisfaction Items

Initially, we checked the factor structure of the JS items by means of an exploratory factor analysis (promax rotation) applied to a polychoric correlation matrix of the items (Jöreskog and Moustaki 2001) using the R package lavaan version 0.6–1 (Rosseel 2012). This analysis revealed that the JS items had an oblique three-dimensional structure in all experimental conditions (see Table S10 in the supplementary material). Table 2 presents the JS subscales and descriptive statistics for the JS items under different experimental conditions. Independent of rating scale length, the respondents were more satisfied with work tasks and working conditions and with the social aspects of their jobs and less satisfied with job-related benefits and prospects. In addition, the ordering of the average satisfaction level of job aspects within a particular subscale did not differ between conditions (with the exception of the first subscale in the 6-category condition). Thus, respondents were most satisfied with job security (in terms of benefits and prospects), working conditions (regarding work tasks and conditions), work atmosphere, and relationships with co-workers (e.g., the social aspects of their jobs) and least satisfied with the non-monetary benefits of their jobs, the work tasks themselves, and the appreciation, recognition, and rewards they received for good work. However, the variance in

**Table 2** Descriptive statistics for the items of job satisfaction and the reliability scores of subscales under the different experimental conditions

Item	11-category condition			6-category condition			4-category condition		
	Mean (SD)	Skew.	Kurt.	Mean (SD)	Skew.	Kurt.	Mean (SD)	Skew.	Kurt.
<i>Subscale: Benefits and prospects</i>									
Item 1. Payment	5.27 (2.46)	-0.30	-0.58	2.59 (1.24)	-0.26	-0.49	1.50 (0.79)	-0.12	-0.44
Item 2. Monetary and non-monetary fringe benefits	5.16 (2.84)	-0.22	-0.93	2.64 (1.46)	-0.25	-0.90	1.49 (0.92)	-0.08	-0.83
Item 3. Job security	6.36 (2.87)	-0.64	-0.57	3.24 (1.47)	-0.69	-0.46	1.89 (0.97)	-0.49	-0.74
Item 4. Career opportunities	5.44 (2.95)	-0.23	-0.96	2.68 (1.51)	-0.21	-0.94	1.51 (0.98)	0.00	-1.00
<i>Subscale: Work tasks and conditions</i>									
Item 5. Work tasks	6.58 (2.52)	-0.68	-0.13	3.31 (1.29)	-0.67	-0.05	1.94 (0.86)	-0.46	-0.46
Item 6. Working hours	6.89 (2.63)	-0.75	-0.24	3.50 (1.34)	-0.80	-0.06	2.05 (0.89)	-0.63	-0.40
Item 7. Work conditions	7.32 (2.38)	-0.95	0.40	3.70 (1.24)	-1.00	0.61	2.21 (0.82)	-0.79	-0.01
Item 8. Work-life balance	6.94 (2.79)	-0.80	-0.23	3.53 (1.42)	-0.82	-0.17	2.09 (0.93)	-0.69	-0.52
<i>Subscale: Social aspects</i>									
Item 9. Relationship with the supervisor	6.89 (2.81)	-0.83	-0.21	3.53 (1.41)	-0.89	0.00	2.11 (0.91)	-0.75	-0.34
Item 10. Appreciation, and rewards for good work	6.02 (2.94)	-0.49	-0.79	3.01 (1.51)	-0.44	-0.78	1.72 (0.96)	-0.24	-0.91
Item 11. Relations with coworkers	7.11 (2.48)	-0.93	0.23	3.62 (1.25)	-0.93	0.43	2.13 (0.84)	-0.71	-0.14
Item 12. Staff rules and regulations in the organization	6.59 (2.60)	-0.66	-0.25	3.25 (1.35)	-0.62	-0.24	1.90 (0.89)	-0.43	-0.56

Notes. SD = standard deviation; Skew. = skewness; Kurt. = kurtosis;  $\omega$  = McDonald's omega  
 The numerical value of the lowest category is always zero



the JS variables increased with an increase in the number of response categories. The difference in rating scale length also affected item distributions, which were slightly more left-skewed in the 11- and 6- category conditions than in the 4-category condition (see also the bar plots in Fig. S1, which can be found in the supplementary material). Under the experimental conditions, the values of reliability (McDonald's  $\omega$ ) for all subscales were acceptable, but they decreased as the rating scale was shortened.

## 6.2 Model Fit

To identify the best-fitting model, the relative model fit coefficients of the multidimensional rmGPCM with one to five latent classes were compared within each of the experimental conditions using the BIC. In the 11- and the 6-category conditions, the three-class model was found to be the best-fitting model as it indicated the lowest BIC value (see Table 3). In addition, under both these conditions, the three-class solution provided clearly interpretable class-specific category use patterns. In the 4-category condition, the two-class model showed the lowest BIC value. However, two latent classes did not exhibit any clearly identifiable category use patterns but included elements of appropriate and inappropriate category use. (For example, both classes properly differentiated between response categories. In addition, the first class preferred the middle categories, which covered the entire meaningful range of the latent continuum, whereas the second class preferred the extreme categories.) Therefore, in accordance with other conditions, the item parameter estimates of the three-class solution were inspected. In contrast to the two-class solution, the three-class solution provided a clear separation between inappropriate and appropriate category use patterns. Hence, we also accepted the three-class model as optimal under this condition.

In addition, Table 3 provides the sum of bivariate residuals, indicating that the three-class models apparently worked well under each condition. The mean assignment probabilities and the estimated reliabilities are sufficiently large for all conditions. The values of these statistics were slightly smaller in the conditions with shorter response formats.

## 6.3 Category Use Patterns in Varied Number of Response Categories

Table 4 provides an overview of the major results. In general, rating scale length was found to affect the number of unordered thresholds and category widths but had little effect on the scale range of the JS items. In particular, as a result of reducing the number of response categories, the proportion of reversals (unordered threshold parameters) over all subscales decreased (e.g., for the 11-point rating scale, 8–88%; for the 6-point rating scale, 0–65%; for the 4-point rating scale, 0–25%). This finding indicates that respondents could use the shorter rating scales more effectively and ignored fewer response categories than when using the long rating scale. Furthermore, the response categories of the long rating scale marked smaller segments of the latent continuum (e.g., 0.6–2.0 logits) than those of the 6- and 4-point rating scales (e.g., 0.7–4.3 logits and 0.2–7.6 logits, respectively). However, simultaneously, the average scale range of the JS items remained roughly the same regardless of rating scale length. This indicates that the respondents could better differentiate between a few broadly defined response categories than between many finely defined categories.

In addition, the three latent classes showed some class-specific differences in category use that are consistent across the experimental conditions. In particular, the second class was characterized by a high number of unordered threshold parameters, especially for the long

**Table 3** Goodness-of-fit statistics for the multidimensional rmGPCM in the experimental conditions

Condition	Model	$n_{par}$	LL	BIC	Total BVR	Mean assignment probability in classes <sup>1</sup>	Model-based reliability estimates for JS subscales <sup>2</sup>
11 categories	1 class	135	-56,025	113,097	18,20		
	2 classes	259	-54,080	110,168	5,64		
	3 classes	383	-53,486	<b>109,941</b>	4,72	.91, .92, .89	.84, .84, .86
	4 classes	507	-53,197	110,323	4,62		
	5 classes	631	-52,953	110,796	5,03		
6 categories	1 class	75	-40,368	81,318	6,51		
	2 classes	139	-39,214	79,508	3,57		
	3 classes	203	-38,937	<b>79,450</b>	3,54	.88, .87, .85	.80, .79, .83
	4 classes	267	-38,768	79,611	1,37		
	5 classes	331	-38,640	79,852	1,89		
4 categories	1 class	51	-30,045	60,485	4,01		
	2 classes	91	-29,349	<b>59,403</b>	2,00	.88, .87	.76, .74, .77
	3 classes	131	-29,222	59,458	1,62	.87, .80, .80	.75, .73, .77
	4 classes	171	-29,122	59,569	1,35		
	5 classes	211	-29,046	59,726	2,70		

Notes.  $n_{par}$  = number of model parameters; LL = Log-Likelihood; BIC = Bayesian information criterion; Total BVR = the sum of the bivariate residuals

<sup>1</sup> The values are reported in the following order: for the large class, the middle-sized class, and the small class

<sup>2</sup> The values are reported in the following order: for the subscale "Satisfaction with benefits and prospects", the subscale "Satisfaction with work tasks and conditions", and the subscale "Satisfaction with social aspects"

Within each experimental condition, the BIC value of the best-fitting model is marked in bold

**Table 4** Number of unordered thresholds, category widths, and scale ranges for the rating scales including 11, 6, and 4 response categories

	11-category condition <sup>1</sup>			6-category condition <sup>1</sup>			4-category condition <sup>1</sup>		
	Benefits and prosp.	Work and conditions	Social aspects	Benefits and prosp.	Work and conditions	Social aspects	Benefits and prosp.	Work and conditions	Social aspects
Number of unordered thresholds per item									
Class 1 <sup>2</sup>	2-4	2-4	2-3	0	0-1	0	0	0	0
Class 2	8-9	8-9	7-8	1-4	3-4	2-3	0-1	0-2	0-1
Class 3	0-2	1-2	0-3	0-1	0	0	0	0	0
Proportion of unordered thresholds in a subscale									
Class 1 <sup>2</sup>	27.5	27.5	22.5	0	5	0	0	0	0
Class 2	87.5	85.0	77.5	45	65	55	8.33	25	25
Class 3	7.5	12.5	17.5	5	0	0	0	0	0
Mean (SD) of category widths in a subscale (in logits)									
Class 1 <sup>2</sup>	0.8 (0.4)	0.6 (0.8)	0.7 (0.3)	1.2 (0.8)	0.9 (0.5)	1.1 (0.7)	2.3 (0.8)	1.6 (0.3)	2.0 (0.5)
Class 2	0.8 (0.3)	1.0 (0.1)	1.8 (0.9)	0.8 (0.6)	0.7 (0.3)	0.9 (0.1)	0.8 (0.6)	0.2 (0.2)	0.2 (0.2)
Class 3	2.0 (0.9)	1.7 (1.7)	1.7 (1.3)	3.1 (1.4)	2.7 (1.4)	4.3 (2.2)	6.8 (2.2)	6.3 (2.0)	7.6 (2.0)
Average scale range and SD in parentheses (in logits)									
Class 1 <sup>2</sup>	4.1 (0.5)	3.8 (1.6)	4.0 (1.1)	4.7 (0.9)	3.4 (0.5)	4.5 (1.0)	4.5 (1.1)	3.3 (0.5)	4.1 (0.7)
Class 2	3.0 (1.3)	2.4 (1.3)	2.4 (0.5)	1.2 (0.6)	1.4 (0.8)	1.2 (0.8)	1.6 (1.0)	0.5 (0.4)	0.4 (0.3)
Class 3	14.4 (4.1)	12.2 (2.8)	11.6 (2.6)	11.7 (1.0)	10.2 (3.0)	17.1 (2.4)	13.5 (3.3)	12.5 (3.0)	15.3 (1.3)

*Notes.* *Benefits and prosp.* = subscale "Satisfaction with benefits and prospects"; *Work and conditions* = subscale "Satisfaction with work tasks and conditions"; *Social aspects* = subscale "Satisfaction with social aspects"; *SD* = standard deviation

<sup>1</sup> An item with 11, 6, and 4 response categories has 10, 5, and 3 thresholds, respectively

<sup>2</sup> Latent classes within each of the experimental conditions are sorted in descending order by their size: Class 1 is the largest, while the Class 3 is the smallest

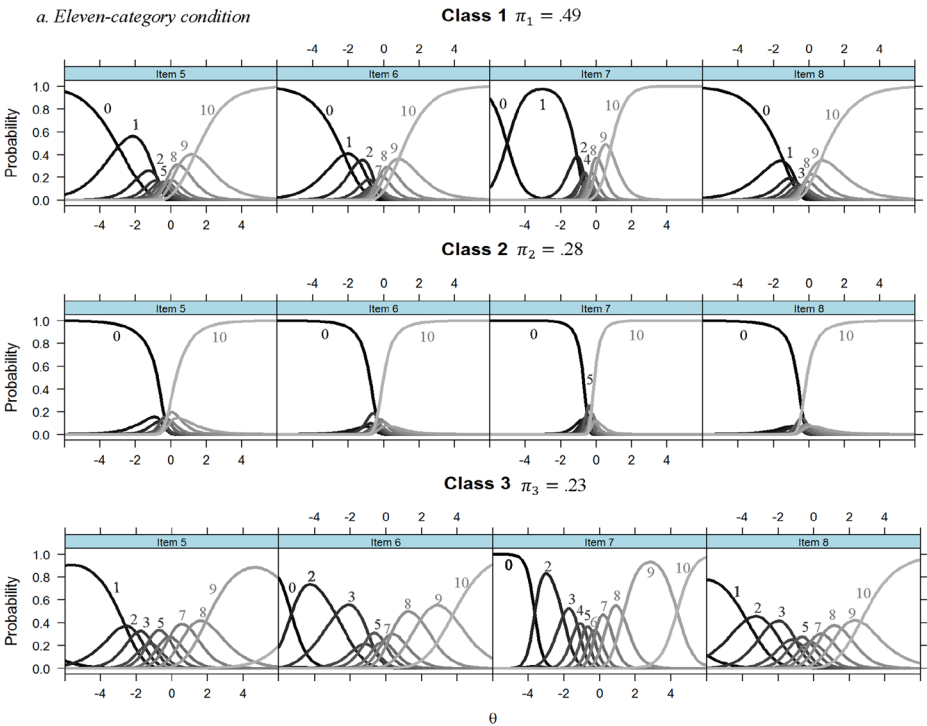
The category width was calculated as the difference between two adjacent thresholds of an item. Scale range of an item was defined as the difference between its highest and lowest thresholds

rating scale (7–9); in the other two classes, few reversals were found (max. 4). This result highlights that classes differ in the number of categories that were actually used. Furthermore, in the third class, response categories marked large segments of the latent continuum (on average, at least 1.7–6.3 logits across all conditions), with the result that an enormously large scale range was covered (e.g., on average, at least 10.2 logits across all conditions) compared to the other two classes. This means that only a few response categories in this class are located inside the reasonable area of the latent continuum.

Next, we explain in detail the results for only one subscale (“Satisfaction with work tasks and conditions”), as the JS subscales within the conditions did not differ much in the general results obtained. Finally, we address some specific features of the category use patterns for the other two JS subscales.

### 6.3.1 Class-Specific Category Use for the “Satisfaction with work tasks and conditions” Subscale

Figure 2 presents the class-specific CCCs for the three different rating scales. (For the estimated item parameters used to draw these CCCs, see Table S11–S13 in the supplementary material.) For the 11-category condition (Fig. 2a), the first class, which included nearly half of the sample (49%), indicated that, depending on the item considered, up to three lower and three upper response categories have areas on the latent variable in which their response probabilities are higher than those of other categories. The CCCs of the middle categories ( $x =$



**Fig. 2** Class-specific category characteristic curves for the items of the subscale “Satisfaction with work tasks and conditions” in the three experimental conditions. (Categories whose response probability is the highest on a certain segment of the latent continuum are indicated with their values)

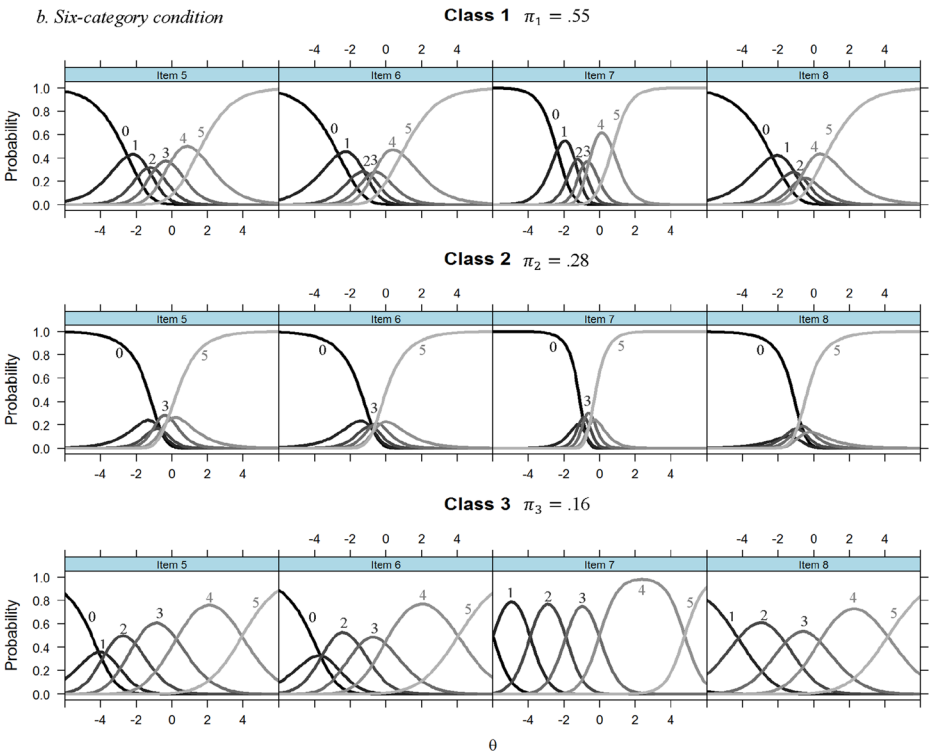


Fig. 2 (continued)

4 to  $x = 7$ ) are close to each other, indicating that these categories are avoided or that only one of them covers a very small segment of the latent continuum. The other half of the sample consisted of two latent classes. The medium-sized second class (28%) was characterized by using only the lowest and the highest response categories; all the categories in between were avoided. In the small third class (23%), the threshold parameters were generally ordered, indicating that this class used the subscale in the intended way. However, the extreme categories in this latent class had the highest response probability outside a meaningful range of the latent continuum (below  $\theta_{vtg} = -4$  and above  $\theta_{vtg} = 4$ ). This means that the latent trait values of both very dissatisfied and very satisfied respondents were unreasonably low or high. Moreover, the respondents belonging to the last class tended to ignore two categories of the middle part of the response format ( $x = 4$  and  $x = 6$ ). Considering the class-specific category use pattern described above, the first class exhibited the ordinary response style (the ORS class), whereas the second class clearly demonstrated an ERS (the ERS class). The third class did not use any common RS. The category use of the members of this class was characterized by avoiding extreme response categories. For that reason, we classified this class as a non-ERS class.

For the 6-category condition (Fig. 2b), latent classes with similar category use patterns were found: the ORS class, the ERS class, and the non-ERS class. The ORS of the large class (55%) was marked by an appropriate distinction between six response categories (except item 3, one of the middle categories, namely  $x = 3$ , of which was ignored). However, compared to other

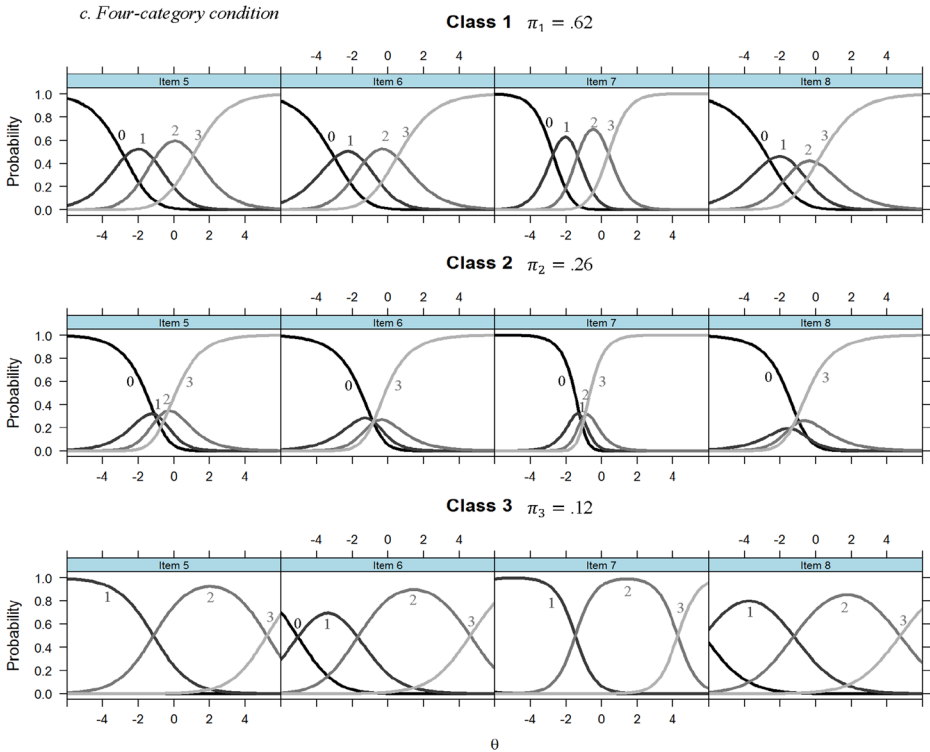


Fig. 2 (continued)

categories, the two middle categories ( $x = 2$  and  $x = 3$ ) covered smaller segments of the latent continuum. The medium-sized class (28%) preferred extreme categories (ERS) and also, for items 5 and 7, one of the middle categories ( $x = 3$ ). The small class (16%) with a non-ERS was characterized by the use of non-extreme categories within a reasonable range of the latent continuum ( $x = 2$  to  $x = 4$ ).

In the 4-category condition (Fig. 2c), the three classes also exhibited similar patterns of category use as those described above. In the ORS class (62%), all four response categories covered equidistant segments on the latent continuum. The respondents who fell into the ERS class (26%) mostly used only two extreme categories. Although four categories were present on the latent continuum in the non-ERS class (12%), only two middle categories ( $x = 1$  and  $x = 2$ ) were related to a meaningful range of the latent trait variable.

### 6.3.2 Specific Features of the Other Job Satisfaction Subscales

By and large, the patterns of class-specific category use within the various experimental conditions were similar across the JS subscales (see Figs. S2–S3 in the supplementary material). However, a few specific features could be identified for the subscale “Satisfaction with benefits and prospects”. For example, the ORS class ignored some categories in the middle segment of the response format ( $x = 4$  to  $x = 6$  for all items in the 11-category condition;  $x = 3$  for item 3 and  $x = 2$  for item 4 in the 6-category condition). In the ERS class, one of the

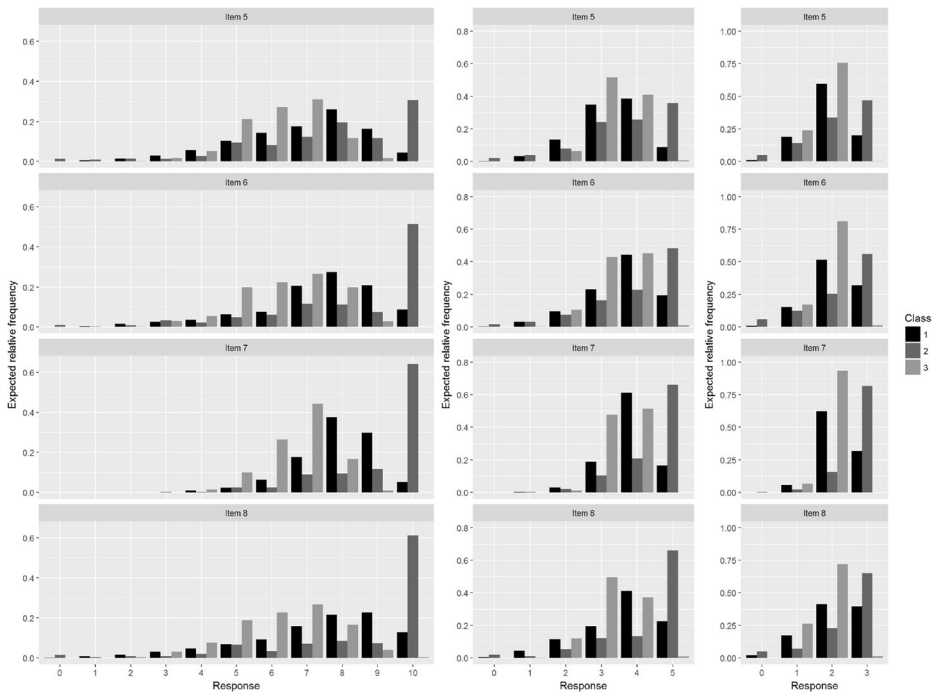
middle categories (if presented) covered a slightly larger segment of the latent continuum, as was observed for the subscale “Satisfaction with work tasks and conditions” (this concerns  $x = 5$  or  $x = 7$ ,  $x = 3$  or  $x = 4$ , and  $x = 1$  or  $x = 2$  in the 11-, 6-, and 4-category conditions, respectively). In the non-ERS class, more categories ( $x = 1$  to  $x = 4$ ) were present within the meaningful range of the latent continuum in the 6-category condition. For the subscale “Satisfaction with social aspects”, specific features were also related to the use of categories in the middle segment of the rating scale. In the ORS class, we found that category  $x = 3$  covered a larger segment of the latent trait variable in the 6-category condition. In the ERS class, one of the middle categories had the highest probability of being endorsed only on a small area (this concerns  $x = 5$  and  $x = 3$  in the 11- and 6-category conditions, respectively).

In summary, the results reported in this section revealed similar patterns of category use in the form of the ORS, ERS, and non-ERS existing in the JS data assessed with different rating scales. However, the proportion of respondents who exhibited ORS increased with a reduction in the number of response categories (for the 11-category condition, 49%; for the 6-category condition, 55%; and, for the 4-category condition, 62%). Coincidentally, the proportion of respondents who demonstrated the non-ERS decreased (for the 11-category condition, 23%; for the 6-category condition, 16%; and, for the 4-category condition, 12%). Beyond that, (almost) all of the categories of the shorter rating scales were actually used by respondents (with the exception of the ERS class), whereas many categories of the long rating scale were ignored. Thus, these results confirmed our first expectation, namely that shorter rating formats trigger less inappropriate category use in comparison with rating scales consisting of many response categories. Furthermore, the ERS was used by almost a third of the sample, regardless of the experimental conditions (for the 11- and 6-category conditions, 28%; and for the 4-category condition, 26%). This finding is consistent with previous findings that have indicated that the ERS is used consistently, regardless of rating scale length.

## 6.4 Expected Relative Category Frequencies of Job Satisfaction Items in Latent Classes

Figure 3 presents the expected relative frequencies of the response categories for the subscale “Satisfaction with work tasks and conditions”. Generally, they depend on the threshold parameters and the distribution of the latent trait variables in the latent classes. For this subscale, lower categories primarily showed low frequencies, despite the different rating scale lengths (e.g.,  $x = 0$  to  $x = 4$ ,  $x = 0$  to  $x = 2$ , and  $x = 0$  for the 11-, 6-, and 4-category conditions, respectively). This means that respondents were generally satisfied with these aspects of their jobs. A more fine-grained analysis indicated that the top category was selected most frequently in the ERS class, regardless of the number of response categories. In contrast, the ORS class exhibited a preference for other upper categories (e.g.,  $x = 7$  to  $x = 9$  for the 11-category condition and  $x = 4$  for the 6-category condition); the non-ERS class showed more frequent use of middle and upper categories (e.g.,  $x = 6$  to  $x = 7$  for the 11-category condition and  $x = 3$  to  $x = 4$  for the 6-category condition). However, in the 4-category condition, both of these classes selected the second-highest category most frequently. This means that the shortest rating scale minimized differences in their class-specific category use.

Similar results were obtained for the subscale “Satisfaction with social aspects”. Most of the items of the subscale “Satisfaction job-related benefits and prospects” were approximately symmetrically distributed, indicating that respondents were less satisfied with this job aspect (Figs. S4–S5 in the supplementary material).



**Fig. 3** Expected relative frequencies for the items of the subscale “Satisfaction with work tasks and conditions” in the three experimental conditions (11-category condition in the left column, 6-category condition in the middle column, and 4-category condition in the right column)

## 6.5 Careless Responding

In the next step, we analyzed the number of insufficient-effort respondents and their distribution over the latent classes. Under all experimental conditions, we found a slightly higher proportion of respondents with random responses (max. 7.1%), followed by individuals who exhibited inattentive responding (max. 6.8%) and those who provided invariant responses (max. 6.3%); quick responding occurred less frequently (max. 3.9%; see Table 5). This number of insufficient-effort respondents is below the modal proportion reported in previous research on this issue (e.g., 8–12%; see Curran 2016 for an overview). Across all experimental conditions, latent classes with identical RSs hardly differed in the distribution of careless responding. However, within each of the experimental conditions, the latent classes exhibited significant differences in careless responding. In particular, in the non-ERS class, respondents were found to provide inattentive or invariant responses and to exhibit quick responding approximately twice as frequently as those in the other two latent classes. An exception was the ERS class, which included fewer respondents who provided invariant responses than the non-ERS class but more than in the ORS class. Conversely, random responses were mostly present in the ERS class, followed by the ORS class (with the exception of the 4-category condition); in the non-ERS class, the proportion of respondents who demonstrated this type of careless responding was at a minimal level. Hence, these results indicate an association between class-specific category use and forms of careless responding.



**Table 5** Proportion of careless responding in latent classes within the experimental conditions (%)

Index	11-category condition (n = 2322)	6-category condition (n = 2364)	4-category condition (n = 2313)	Test statistics
Inattentive responding	6.8	6.2	6.2	Between conditions: $\chi^2(2) = 0.76, p = .68$
ORS class	5.7	5.4	5.1	Between classes within: 4-category condition: $\chi^2(2) = 28.59, p < .001$
ERS class	5.3	4.5	5.6	6-category condition: $\chi^2(2) = 25.47, p < .001$
Non-ERS class	10.7	11.7	13.4	11-category condition: $\chi^2(2) = 17.65, p < .001$
Quick responding	3.3	2.8	3.9	Between conditions: $\chi^2(2) = 4.43, p = .11$
ORS class	2.4	1.8	3.1	Between classes within: 4-category condition: $\chi^2(2) = 38.77, p < .001$
ERS class	1.5	3.3	2.6	6-category condition: $\chi^2(2) = 13.81, p < .01$
Non-ERS class	7.2	5.2	10.6	11-category condition: $\chi^2(2) = 35.42, p < .001$
Invariant responding	5.7	6.3	5.2	Between conditions: $\chi^2(2) = 2.49, p = .29$
ORS class	3.6	4.6	3.2	Between classes within: 4-category condition: $\chi^2(2) = 39.13, p < .001$
ERS class	6.5	6.0	7.4	6-category condition: $\chi^2(2) = 28.08, p < .001$
Non-ERS class	9.2	11.9	11.3	11-category condition: $\chi^2(2) = 22.44, p < .001$
Random responding	7.1	6.5	6.8	Between conditions: $\chi^2(2) = 0.65, p = .72$
ORS class	7.8	5.4	7.3	Between classes within: 4-category condition: $\chi^2(2) = 6.71, p < .05$
ERS class	9.1	9.8	7.2	6-category condition: $\chi^2(2) = 16.03, p < .001$
Non-ERS class	3.0	4.7	3.2	11-category condition: $\chi^2(2) = 19.12, p < .001$

Notes. ORS class = latent class with the ordinary response style; ERS class = latent class with the extreme response style; Non-ERS class = latent class with the non-extreme response style

## 6.6 Prediction of Assignment to Latent Classes

Table 6 presents the results of the multinomial logistic regressions conducted under each of the experimental conditions. In addition to predictors such as respondent characteristics and contextual factors, which were described in the Measures section, we included indices of careless responding; this was done to control for their effects, as we found class-specific differences in these indices under all conditions. For the 4-category condition only, latent trait variables of both satisfaction with work tasks and conditions and satisfaction with social aspects were also included due to the existence of class-specific differences in these JS subscales under this condition (for details concerning the class comparison, see Table S14 in the supplementary material). For reasons of comprehensibility, Table 6 presents only significant predictors and test statistics.<sup>1</sup> Independent of the rating scale, assignment to the ERS class was generally more likely for respondents with higher general self-efficacy and perceived job autonomy. In addition, the probability of being assigned to the ERS class, as opposed to the ORS class, was higher for administrative employees and people with higher scores on self-deceptive enhancement (with the exception of the 4-category condition), as well as for employees in low-level job positions and respondents who provided invariant responses (with the exception of the 6-category condition). This probability became less likely for individuals who reported high job stress (with the exception of the 11-category condition). Furthermore, assignment to the ERS class, as opposed to the non-ERS class, could be predicted based on the presence of random responses and low job stress. It became more likely for self-deceptive respondents and those with high scores on neuroticism (with the exception of the 4-category condition) or higher need for cognition (with the exception of the 6-category condition); however, this type of category use would rarely be practiced by employees in middle-sized organizations (with the exception of the 11-category condition). The non-ERS class could be differentiated from the ORS class by the presence of invariant responses. Beyond the presence of such responses, the probability of being assigned to the non-ERS class increased when a respondent indicated a low level of perceived job security and exhibited an absence of random responses and a low need for cognition. In Table 6, these predictors are marked in gray.

In addition to the predictors reported thus far, an additional set of predictors had an influence under specific experimental conditions (see for non-marked predictors in Table 6). For example, assignment to the ERS class, as opposed to the ORS class, in the 11-category condition was more likely for female employees and respondents with higher impression management and lower agreeableness but less likely for individuals working in middle-sized organizations. In the 6-category condition, assignment to the ERS class, as opposed to the ORS class, could be predicted based on the tendency to provide quick or random responses and a high educational level, a higher level of conscientiousness, as well as working part-time, in a mid-level position, and in a small organization. For the 4-category condition, the ERS was found to be more likely for individuals with lower scores for the verbal analogy task and higher satisfaction with the social aspects of their jobs but less likely for job beginners. The probability of being assigned to the ERS class, as opposed to the non-ERS class, was higher for female employees, indecisive respondents, people working in low-level positions or in a small organization, and those who reported a higher level of job security or exhibited higher impression management (for the 11-category condition). This was also the case for quick responders, part-time employees, respondents with a lower level of job skills, and those who

<sup>1</sup> All results can be obtained from the first author.

**Table 6** Prediction of assignment to latent classes by means of multinomial regression models in the three experimental conditions

	11-category condition				6-category condition				4-category condition			
	<i>B</i>	(SE)	<i>e<sup>B</sup></i>	95%-CI	<i>B</i>	(SE)	<i>e<sup>B</sup></i>	95%-CI	<i>B</i>	(SE)	<i>e<sup>B</sup></i>	95%-CI
<b>ERS class vs. ORS class (reference class)</b>												
Constant	-5.05***	1.14			-5.65***	1.36			-4.10**	1.57		
Quick responding					1.41*	0.57	4.10	[1.34; 12.54]				
Invariant responding	0.73*	0.31	2.07	[1.13; 3.80]					1.13**	0.41	3.09	[1.38; 6.89]
Random responding					0.77*	0.30	2.16	[1.20; 3.89]				
Self-deceptive enhancement	0.23**	0.08	1.26	[1.08; 1.47]	0.27*	0.09	1.31	[1.08; 1.57]				
Impression management	0.18**	0.06	1.20	[1.07; 1.34]								
Gender (female)	0.37**	0.14	1.45	[1.10; 1.91]								
Educational level (high school)					0.77*	0.34	2.16	[1.11; 4.20]				
Conscientiousness					0.24*	0.11	1.27	[1.03; 1.57]				
Agreeableness	-0.16*	0.08	0.85	[0.73; 0.99]								
General self-efficacy	0.70**	0.21	2.01	[1.32; 3.06]	0.59*	0.28	1.81	[1.06; 3.11]	0.83**	0.28	2.29	[1.33; 3.93]
Verbal memory ability	-0.08**	0.03	0.92	[0.87; 0.98]								
Verbal analogy task												
Job position (Level 3)					0.53*	0.26	1.70	[1.02; 2.84]				
Job position (Level 4)	0.48*	0.22	1.62	[1.04; 2.52]	0.56*	0.27	1.76	[1.04; 2.95]				
Job position (Level 5)	0.68**	0.22	1.97	[1.27; 3.06]					0.58*	0.29	1.79	[1.02; 3.14]
Tenure at position (1-3 years)									-0.71*	0.31	0.49	[0.27; 0.91]
Part-time occupation					0.55**	0.18	1.73	[1.23; 2.44]				
Organization size (small)					0.35*	0.18	1.42	[1.01; 2.00]				
Organization size (middle)	-0.43*	0.18	0.65	[0.46; 0.92]								
Job autonomy	0.21***	0.06	1.23	[1.10; 1.38]	0.19**	0.07	1.21	[1.06; 1.38]	0.23**	0.09	1.26	[1.07; 1.49]
Job-related stress					-0.12*	0.06	0.89	[0.79; 0.99]	-0.19*	0.08	0.83	[0.71; 0.98]
Satisfaction with social aspects									0.30*	0.12	1.35	[1.07; 1.70]
<b>ERS class vs. non-ERS class (reference class)</b>												
Constant	-7.37***	1.24			-7.11***	1.39			-4.73**	1.73		
Quick responding					0.92*	0.43	2.50	[1.08; 5.82]				
Invariant responding					-0.80**	0.31	0.45	[0.25; 0.82]				
Random responding	1.65***	0.44	5.22	[2.22; 12.28]	0.95*	0.39	2.58	[1.21; 5.51]	1.82**	0.70	6.16	[1.55; 24.41]
Self-deceptive enhancement	0.29***	0.08	1.34	[1.14; 1.58]	0.27**	0.10	1.31	[1.09; 1.58]				
Impression management	0.17**	0.06	1.19	[1.05; 1.35]								
Gender (female)	0.49**	0.16	1.63	[1.20; 2.23]								
Neuroticism	0.24**	0.09	1.28	[1.07; 1.53]	0.33**	0.10	1.39	[1.13; 1.70]				
Openness to experience					0.20*	0.10	1.23	[1.01; 1.48]				
General self-efficacy	0.93***	0.23	2.53	[1.62; 3.96]	1.36***	0.28	3.91	[2.24; 6.83]	1.15***	0.32	3.17	[1.70; 5.92]
Need for cognition	0.24**	0.12	1.28	[1.01; 1.61]					0.38*	0.17	1.47	[1.05; 2.04]
Decisiveness	-0.24*	0.11	0.78	[0.63; 0.97]								
Tolerance for ambiguity												
Job position (Level 5)	0.51*	0.25	1.66	[1.02; 2.71]					-0.39*	0.16	0.68	[0.49; 0.94]
Part-time occupation					0.56**	0.20	1.75	[1.18; 2.59]				
Organization size (small)												
Organization size (middle)	0.51**	0.18	1.67	[1.18; 2.35]	-0.72**	0.23	0.49	[0.31; 0.76]	-0.71**	0.26	0.49	[0.30; 0.83]
Job autonomy	0.17**	0.06	1.19	[1.06; 1.33]	0.21**	0.07	1.23	[1.08; 1.41]	0.31**	0.09	1.36	[1.13; 1.63]
Job skills					-0.18**	0.07	0.83	[0.73; 0.95]				
Job-related stress	-0.22***	0.05	0.80	[0.73; 0.89]	-0.13*	0.06	0.88	[0.78; 0.98]	-0.25**	0.08	0.78	[0.67; 0.91]
Job security	0.16*	0.06	1.17	[1.03; 1.33]								
<b>Non-ERS class vs. ORS class (reference class)</b>												
Constant	2.32*	1.01										
Inattentive responding												
Invariant responding	0.86**	0.28	2.37	[1.37; 4.11]	0.87**	0.29	2.38	[1.35; 4.19]	0.75*	0.32	2.11	[1.12; 3.99]
Random responding	-1.38**	0.42	0.25	[0.11; 0.57]					1.58***	0.38	4.84	[2.31; 10.14]
Neuroticism					-0.19*	0.09	0.83	[0.70; 0.98]	-1.61*	0.69	0.20	[0.05; 0.78]
General self-efficacy					-0.77***	0.21	0.46	[0.31; 0.70]				
Need for cognition	-0.32**	0.10	0.72	[0.59; 0.88]					-0.30*	0.12	0.74	[0.59; 0.94]
Decisiveness					0.26*	0.12	1.30	[1.03; 1.64]				
Tenure at position (4-6 years)	0.80**	0.27	2.22	[1.31; 3.78]								
Organization size (middle)					0.52**	0.20	1.68	[1.14; 2.47]				
Job-related stress	0.13***	0.04	1.14	[1.06; 1.23]								
Job security	-0.12*	0.05	0.89	[0.81; 0.98]	-0.12*	0.05	0.88	[0.80; 0.98]				
Quasi- <i>R</i> <sup>2</sup> (Nagelkerke)		.18				.15				.15		
Fit improvement of the model compared to the baseline model				$\chi^2(76) = 392.67, p < .001$				$\chi^2(76) = 334.60, p < .001$				$\chi^2(80) = 301.18, p < .001$

Notes. *B* = estimated regression coefficient; *SE* = standard error of an estimated regression coefficient; *e<sup>B</sup>* = exponential of *B* (odds ratio); 95%-*CI* = 95% confidence interval of the odds ratio. *ERS class* = latent class with the extreme response style; *ORS class* = latent class with the ordinary response style; *non-ERS class* = latent class with the non-extreme response style

Reference categories of nominal predictors: For educational level (1 = high school graduate or less), for job position: (level 1: manager, self-employed, etc.), for tenure at current position (5 = 10 years or longer), and for organization size (large: 200 employees or more)

Predictors which had a significant effect on predicting class assignment at least for two types of rating scale lengths are highlighted in gray

\* *p* < .05, \*\* *p* < .01, \*\*\* *p* < .001

were open to new experiences. It was less likely for respondents with invariant responses (for the 6-category condition). In the 4-category condition, both latent classes could be additionally separated by the low tolerance for ambiguity that was more likely to be found among those

who belonged to the ERS class. Finally, assignment to the non-ERS class, as opposed to the ORS class, was more likely for long-tenure workers, respondents with higher levels of job stress (for the 11-category condition), those who were decisive, those with lower scores on neuroticism or low general self-efficacy, those working in middle-sized organizations (for the 6-category condition), and those who were inattentive (for the 4-category condition). For any proposed rating scale, the effects of age, job relevance, and extraversion could not be distinguished among the latent classes by controlling for other predictors.

## 7 Discussion

In this paper, we used a randomized experimental design to investigate whether short rating scales (with 4 or 6 categories) would produce a lower extent of inappropriate category use, compared to a long rating scale (which consists of 11 categories), when assessing JS. Using the multidimensional rmGPCM, we found similar category use patterns, namely the ORS (ORS class), the ERS (ERS class), and the non-extreme RS (non-ERS class), under all experimental conditions. This similarity of category use patterns existing in the data regardless of rating scale length is a new finding. It was obtained by exploring RSs using the mixed IRT model, which does not require an a priori definition of any RSs. It follows that RSs are not only a methodological nuisance but should also be considered a trait of substantive meaning (Kieruj and Moors 2013). In particular, this refers to the ERS, which, as our data indicated, was used by almost one-third of the respondents, regardless of rating scale length (26–28%). This finding is in line with previous research that has found that the ERS is an individual response style that occurs across measures and points of measurement (e.g., Kieruj and Moors 2013; Weijters et al. 2010a, 2010b; Wetzel et al. 2013; Wetzel et al. 2016; Zettler et al. 2016). As a result, our study seems to indicate that some RSs can hardly be eliminated by shortening rating scale length. This implies that survey practitioners should apply statistical methods to control for these consistent RS effects.

However, other relevant results obtained in the present study confirmed our hypothesis that short rating scales lead to the reduced presence of RSs, suggesting that the respondents could effectively cope with a less demanding rating scale. It was found that the proportion of respondents exhibiting inappropriate category use (the non-ERS class) decreased with a reduction in the number of response categories (from 23% to 16% and 12% in the 11-, 6-, and 4-category conditions, respectively). Coincidentally, the proportion of respondents using the rating scale in an ordinary way (the ORS class) increased (from 49% to 55% and 62% in the 11-, 6-, and 4-category conditions, respectively). Thus, shortening a rating scale may reduce the number of respondents using RSs as an adjustment strategy to overcome difficulties they experience in handling the long rating scale. This finding is in accordance with Krosnick's (1991) concept of two types of response behavior: optimizing, which occurs when respondents endeavor to respond appropriately, and satisficing, which is characterized by taking cognitive shortcuts and employing adjustment strategies due to suboptimal rating scale features. In addition, it was found that the short rating scales showed almost no unordered thresholds, indicating that respondents actually used (almost) all categories of these response formats (with the exception of the ERS class). In the long rating scale, many response categories were ignored. As such, offering short rating scales may avoid overloading the respondents' differentiation ability. This finding is also consistent with the recommendation provided by previous research that for the general population, rating scale length should not exceed six or seven

response categories (e.g., Lozano et al. 2008). This conclusion can be supported by the findings on reliability. It was found to be high for the 6-point rating scale. However, for the 11-point rating scale, reliability was slightly higher; for the 4-point rating scale, it was slightly lower, but still sufficiently high. When the number of response categories decreases a general decrease in reliability may in part emerge due to skewed item distributions and lower total score variability (Bandalos and Enders 1996; Masters 1974). Both points hold for our data; the JS items were skewed distributed in all conditions and showed smaller variances when a rating scale included a few response categories. Nevertheless, by optimizing a rating scale by means of including only four or six response categories instead of eleven, researchers may be able to prevent the use of RSs as a form of adjustment strategy and thus obtain high-quality data.

Because the use of RSs can also be caused by stable respondent characteristics and because previous empirical findings concerning this issue did not provide any systematic knowledge, the second aim of this study was to conduct a systematic examination of which respondent characteristics and job-related factors would consistently explain the RSs that were found when JS was measured with different rating scales. An important result of the present study is that we found two sets of predictors: (i) the so-called (almost) general predictors, which showed a statistically significant effect on predicting the use of a specific RS, (almost) regardless of the rating scale length used to assess JS; and (ii) the so-called response-format-specific predictors, which showed a statistically significant effect under only one experimental condition. In particular, the ERS users under all experimental conditions were characterized by a high level of general self-efficacy and perceived job autonomy. This personality profile can be complemented by the almost general predictors, which could account for the use of a certain RS under two experimental conditions. Compared to ORS users, the ERS users also worked in low or mid-level job positions and reported high levels of self-deception and low levels of job-related stress. They were also inclined to invariant responding. Compared to non-ERS users, ERS users were more likely to have high levels of self-deception, neuroticism, and need for cognition and low levels of job-related stress. They also showed random responding. For non-ERS users, a tendency to invariant responding (a general predictor) as well as a low level of job security and low need for cognition (almost general predictors) were substantial characteristics compared to the respondents who engaged in the ORS. The response-format-specific predictors included socio-demographic variables (gender and education level), personality traits (impression management, agreeableness, conscientiousness, openness to experience, decisiveness, and tolerance for ambiguity), cognitive ability, types of careless responding and the majority of job-related factors (organization size, part-time working, job skills, and tenure in current position). This finding indicates that the characteristics of respondents using a certain RS as an adjustment strategy may differ depending on rating scale length, whereas the respondents who use RSs due to stable individual dispositions have consistent personality profiles found across the different rating scales. Age, job relevance, and extraversion were found to be statistically significant in none of the experimental conditions.

## 7.1 Limitation and Future Research

The generalizability of the reported results is limited due to the experimental design, in which only the number of response categories was varied, with other features of the rating scale being set to be equal across the experimental conditions. It is known from previous research that the effect of rating scale length on category use and reliability may be strengthened or mitigated by additional features of a rating scale that may provide respondents with additional cues for

interpretation (Caboooter et al. 2016; Tourangeau et al. 2007). For example, compared to fully labeled rating scales, endpoint-labeled rating scales may prove more challenging for respondents because the meaning of intermediate categories may remain unclear (Hamby and Levine 2016). For this reason, endpoint-labeled rating scales should be shorter to avoid potential cognitive overload and a higher risk of respondents engaging in the ERS (Moors et al. 2014; Weijters et al. 2010). Furthermore, respondents perceive unipolar rating scales with positively and negatively numbered categories to be rather symmetrical compared to those with only positively numbered categories. Therefore, the former may prompt a lower extent of RSs (Caboooter et al. 2016; Moors et al. 2014). Thus, we can primarily generalize our findings to rating scales with both endpoint-labeled and positively numbered response categories.

A further limitation arises from the confounding of rating scale length with the inclusion or omission of the middle category due to the experimental design: an odd-numbered rating scale (11-category condition) was compared with even-numbered rating scales (4- and 6-category conditions). Therefore, the presence of the middle category in the 11-category condition may additionally strengthen a potential effect of rating scale length on RS use (see Kieruj and Moors 2010; Moors 2008; O'Muircheartaigh et al. 1999; Weijters et al. 2010). Moreover, the findings of this study essentially hold for JS. In addition, they could be reasonably generalized to other aspects of cognitive well-being (e.g., satisfaction with family life, health, and home). However, the use of RSs may be partly trait-specific (for ERS, see Caboooter et al. 2017). Due to this specificity of RSs, the generalizability of these findings to other traits (e.g., personality traits) is limited. Future research may replicate the findings of this study for rating scales with other features (e.g., fully labeled) and other constructs (e.g., personality traits).

In general, this experimental study addressed two criteria for evaluating the optimality of rating scale length: the occurrence of inappropriate category use and reliability. Overall, the evidence obtained in this study highlights the superiority of the short rating scales (namely 4- and 6-point rating scales) over the long one (namely an 11-point rating scale) in terms of the extent of inappropriate category use. However, the optimality of a rating scale is a broad concept which includes further practical aspects (e.g., ease of use, the time required for responding) and types of validity (e.g., predictive validity). Future research should consider these further aspects to provide strong support for the optimality of the short rating scales.

A further suggestion for future research is to examine the antecedents of inappropriate category use. The present study examined the effects of rating scales and several individual variables on the use of RSs. Therefore, these findings are limited due to the variables considered. In addition, the majority of respondent characteristics, which were used in this study to predict RSs, were measured using rating scales and respondents' responses to these scales may therefore have been affected by RSs. Researchers could examine to what extent inappropriate responses are related to additional individual variables (e.g., motivation to participate, attitude towards accurate responding, knowledge of the topic, and relevance of the topic, as well as mood, fatigue, and level of concentration). Moreover, researchers could investigate the interactions of individual variables with item characteristics as a further potential antecedent of RSs. For example, respondents may demonstrate a higher tendency to engage in the ERS in response to items that they consider more important.

**Funding** Open Access funding enabled and organized by Projekt DEAL.

## Compliance with Ethical Standards

**Conflict of Interest** The authors declare that there is no conflict of interest.

**Ethical Approval** All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

**Informed Consent** Informed consent was obtained from all individual participants included in the study.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Alwin, D. F., & Krosnick, J. A. (1991). The reliability of survey attitude measurement: The influence of question and respondent attributes. *Sociological Methods and Research*, 20(1), 139–181. <https://doi.org/10.1177/0049124191020001005>.
- Andrich, D. (2010). Understanding the response structure and process in the polytomous Rasch model. In M. L. Nering & R. Ostins (Eds.), *Handbook of polytomous item response models* (pp. 123–152). New York: Routledge.
- Austin, E. J., Deary, I. J., & Egan, V. (2006). Individual differences in response scale use: Mixed Rasch modelling of responses to NEO-FFI items. *Personality and Individual Differences*, 40(6), 1235–1245. <https://doi.org/10.1016/j.paid.2005.10.018>.
- Bandalos, D. L., & Enders, C. K. (1996). The effects of nonnormality and number of response categories on reliability. *Applied Measurement in Education*, 9, 151–160. [https://doi.org/10.1207/s15324818ame0902\\_4](https://doi.org/10.1207/s15324818ame0902_4).
- Baumgartner, H., & Steenkamp, J.-B. E. M. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research*, 38(2), 143–156. <https://doi.org/10.1509/jmkr.38.2.143.18840>.
- Beauducel, A. (2010). *Intelligence structure test: IST; English version of the "Intelligenz-Struktur-Test 2000 R (IST 2000 R)"* by D. Liepmann, A. Beauducel, B. Brocke & R. Amthauer; manual. Hogrefe.
- Billiet, J. B., & Davidov, E. (2008). Testing the stability of an acquiescence style factor behind two interrelated substantive variables in a panel design. *Sociological Methods & Research*, 36(4), 542–562. <https://doi.org/10.1177/0049124107313901>.
- Borgers, N., Hox, J., & Sikkels, D. (2004). Response effects in surveys on children and adolescents: The effect of number of response options, negative wording, and neutral mid-point. *Quality & Quantity*, 38, 17–33. <https://doi.org/10.1023/B:QUQU.0000013236.29205.a6>.
- Bowling, N. A., Eschleman, K. J., & Wang, Q. (2010). A meta-analytic examination of the relationship between job satisfaction and subjective well-being. *Journal of Occupational and Organizational Psychology*, 83(4), 915–934. <https://doi.org/10.1348/096317909x478557>.
- Bozdogan, H. (1994). Mixture-model cluster analysis using model selection criteria and a new informational measure of complexity. In H. Bozdogan (Ed.), *Proceedings of the first US/Japan conference on the Frontiers of statistical modeling, vol. 2: An informational approach* (pp. 69–113). Boston: Kluwer Academic Publishers.
- Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, 6(1), 3–5. <https://doi.org/10.1177/1745691610393980>.
- Cabooter, E., Weijters, B., Geuens, M., & Vermeir, I. (2016). Scale format effects on response option interpretation and use. *Journal of Business Research*, 69(7), 2574–2584. <https://doi.org/10.1016/j.jbusres.2015.10.138>.

- Cabooter, E., Weijters, B., De Beuckelaer, A., & Davidov, E. (2017). Is extreme response style domain specific? Findings from two studies in four countries. *Quality & Quantity*, *51*(6), 2605–2622. <https://doi.org/10.1007/s11135-016-0411-5>.
- Cacioppo, J. T., Petty, R. E., Feinstein, J. A., & Jarvis, W. B. G. (1996). Dispositional differences in cognitive motivation: The life and times of individuals varying in need for cognition. *Psychological Bulletin*, *119*(2), 197–253.
- Chang, L. (1994). A psychometric evaluation of 4-point and 6-point Likert-type scales in relation to reliability and validity. *Applied Psychological Measurement*, *18*(3), 205–215. <https://doi.org/10.1177/014662169401800302>.
- Cho, Y. (2013). The mixed distribution Polytomous Rasch model used to account for response styles on rating scales: A simulation study of parameter recovery and classification accuracy. Dissertation, University of Maryland, College Park, MD.
- Churchill Jr., G. A., & Peter, J. P. (1984). Research design effects on the reliability of rating scales: A meta-analysis. *Journal of Marketing Research*, *21*(4), 360–375.
- Clarke, I. (2000a). Extreme response style in cross-cultural research: An empirical investigation. *Journal of Social Behavior and Personality*, *15*(1), 137–152.
- Clarke, I. (2000b). Global marketing research: Is extreme response style influencing your results? *Journal of International Consumer Marketing*, *12*(4), 91–111. [https://doi.org/10.1300/J046v12n04\\_06](https://doi.org/10.1300/J046v12n04_06).
- Cox, E. P. (1980). The optimal number of response alternatives for a scale: A review. *Journal of Marketing Research*, *17*(4), 407–422. <https://doi.org/10.2307/3150495>.
- Culpepper, S. A. (2013). The reliability and precision of total scores and IRT estimates as a function of polytomous IRT parameters and latent trait distribution. *Applied Psychological Measurement*, *37*(3), 201–225. <https://doi.org/10.1177/0146621612470210>.
- Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, *66*, 4–19. <https://doi.org/10.1016/j.jesp.2015.07.006>.
- De Jong, M. G., Steenkamp, J.-B. E. M., Fox, J.-P., & Baumgartner, H. (2008). Using item response theory to measure extreme response style in marketing research: A global investigation. *Journal of Marketing Research*, *45*(1), 104–115. <https://doi.org/10.1509/jmkr.45.1.104>.
- Diener, E., & Suh, E. (1997). Measuring quality of life: Economic, social, and subjective indicators. *Social Indicators Research*, *40*(1–2), 189–216.
- Dziak, J. J., Coffman, D. L., Lanza, S. T., & Li, R. (2012). *Sensitivity and specificity of information criteria* (tech. Rep. No. 12–119). University Park, PA: The Pennsylvania State University, the methodology center. Available from <https://methodology.psu.edu/media/techreports/12-119.pdf>.
- Eid, M., & Rauber, M. (2000). Detecting measurement invariance in organizational surveys. *European Journal of Psychological Assessment*, *16*(1), 20–30. <https://doi.org/10.1027//1015-5759.16.1.20>.
- Freund, P. A., Tietjens, M., & Strauss, B. (2013). Using rating scales for the assessment of physical self-concept: Why the number of response categories matters. *Measurement in Physical Education and Exercise Science*, *17*(4), 249–263. <https://doi.org/10.1080/1091367X.2013.807265>.
- Gerber-Braun, B. (2010). *The double cross: Individual differences between respondents with different response sets and styles on questionnaires*. München: Dissertation, Ludwig-Maximilians-Universität.
- Greenleaf, E. A. (1992a). Improving rating scale measures by detecting and correcting bias components in some response styles. *Journal of Marketing Research*, *29*(2), 176–188. <https://doi.org/10.2307/3172568>.
- Greenleaf, E. A. (1992b). Measuring extreme response style. *Public Opinion Quarterly*, *56*(3), 328–351. <https://doi.org/10.1086/269326>.
- Hamby, T., & Levine, D. S. (2016). Response-scale formats and psychological distances between categories. *Applied Psychological Measurement*, *40*(1), 73–75. <https://doi.org/10.1177/0146621615597961>.
- Harzing, A. W., Baldeuza, J., Barner-Rasmussen, W., Barzantny, C., Canabal, A., Davila, A., et al. (2009). Rating versus ranking: What is the best way to reduce response and language bias in cross-national research? *International Business Review*, *18*(4), 417–432. <https://doi.org/10.1016/j.ibusrev.2009.03.001>.
- Hernández, A., Drasgow, F., & González-Romá, V. (2004). Investigating the functioning of a middle category by means of a mixed-measurement model. *Journal of Applied Psychology*, *89*(4), 687–699. <https://doi.org/10.1037/0021-9010.89.4.687>.
- Huang, H. Y. (2016). Mixture random-effect IRT models for controlling extreme response style on rating scales. *Frontiers in Psychology*, *7*. <https://doi.org/10.3389/fpsyg.2016.01706>.
- Hui, C. H., & Triandis, H. C. (1989). Effects of culture and response format on extreme response style. *Journal of Cross-Cultural Psychology*, *20*(3), 296–309. <https://doi.org/10.1177/0022022189203004>.
- Jin, K. Y., & Wang, W. C. (2014). Generalized IRT models for extreme response style. *Educational and Psychological Measurement*, *74*(1), 116–138. <https://doi.org/10.1177/0013164413498876>.
- Jöreskog, K. G., & Moustaki, I. (2001). Factor analysis of ordinal variables: A comparison of three approaches. *Multivariate Behavioral Research*, *36*(3), 347–387. <https://doi.org/10.1207/S15327906347-387>.



- Judge, T. A., Thoresen, C. J., Bono, J. E., & Patton, G. K. (2001). The job satisfaction–job performance relationship: A qualitative and quantitative review. *Psychological Bulletin*, *127*(3), 376–407. <https://doi.org/10.1037/0033-2909.127.3.376>.
- Khadka, J., Gothwal, V. K., McAlinden, C., Lamoureux, E. L., & Pesudovs, K. (2012). The importance of rating scales in measuring patient-reported outcomes. *Health and Quality of Life Outcomes*, *10*(1), 80–92. <https://doi.org/10.1186/1477-7525-10-80>.
- Kieruj, N. D., & Moors, G. (2010). Variations in response style behavior by response scale format in attitude research. *International journal of public opinion research*, *22*(3), 320–342. <https://doi.org/10.1093/ijpor/edq001>.
- Kieruj, N. D., & Moors, G. (2013). Response style behavior: question format dependent or personal style? *Quality & Quantity*, *47*(1), 193–211. <https://doi.org/10.1007/s11135-011-9511-4>.
- Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*(3), 213–236. <https://doi.org/10.1002/acp.2350050305>.
- Kulas, J. T., & Stachowski, A. A. (2009). Middle category endorsement in odd-numbered Likert response scales: Associated item characteristics, cognitive demands, and preferred meanings. *Journal of Research in Personality*, *43*(3), 489–493. <https://doi.org/10.1016/j.jrp.2008.12.005>.
- Kulas, J. T., & Stachowski, A. A. (2013). Respondent rationale for neither agreeing nor disagreeing: Person and item contributors to middle category endorsement intent on Likert personality indicators. *Journal of Research in Personality*, *47*(4), 254–262. <https://doi.org/10.1016/j.jrp.2013.01.014>.
- Kutscher, T., Crayen, C., & Eid, M. (2017). Using a mixed IRT model to assess the scale usage in the measurement of job satisfaction. *Frontiers in Psychology*, *7*, 1998. <https://doi.org/10.3389/fpsyg>.
- Kutscher, T., Eid, M., & Crayen, C. (2019). Sample size requirements for applying mixed Polytomous item response models: Results of a Monte Carlo simulation study. *Frontiers in Psychology*, *10*, 2494. <https://doi.org/10.3389/fpsyg.2019.02494>.
- Lee, J., & Paek, I. (2014). In search of the optimal number of response categories in a rating scale. *Journal of Psychoeducational Assessment*, *32*(7), 663–673. <https://doi.org/10.1177/0734282914522200>.
- Lozano, L. M., García-Cueto, E., & Muñiz, J. (2008). Effect of the number of response categories on the reliability and validity of rating scales. *Methodology*, *4*(2), 73–79.
- MacDonald Jr., A. P. (1970). Revised scale for ambiguity tolerance: Reliability and validity. *Psychological Reports*, *26*(3), 791–798.
- Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon’s mechanical Turk. *Behavior Research Methods*, *44*(1), 1–23. <https://doi.org/10.3758/s13428-011-0124-6>.
- Masters, J. R. (1974). The relationship between number of response categories and reliability of Likert-type questionnaires. *Journal of Educational Measurement*, *11*, 49–53. <https://doi.org/10.1111/j.1745-3984.1974.tb00970.x>.
- Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models. *Measurement*, *11*(3), 71–101. <https://doi.org/10.1080/15366367.2013.831680>.
- Maydeu-Olivares, A., & Joe, H. (2008). An overview of limited information goodness-of-fit testing in multidimensional contingency tables. In K. Shigemasa, A. Okada, T. Imaizumi, & T. Hoshino (Eds.), *New trends in psychometrics* (pp. 253–262). Tokyo: Universal Academy Press.
- Maydeu-Olivares, A., Kramp, U., García-Forero, C., Gallardo-Pujol, D., & Coffman, D. (2009). The effect of varying the number of response alternatives in rating scales: Experimental evidence from intra-individual effects. *Behavior Research Methods*, *41*(1), 295–308. <https://doi.org/10.3758/BRM.41.2.295>.
- Meiser, T., & Machunsky, M. (2008). The personal structure of personal need for structure. *European Journal of Psychological Assessment*, *24*(1), 27–34. <https://doi.org/10.1027/1015-5759.24.1.27>.
- Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81–97. <https://doi.org/10.1037/h0043158>.
- Moors, G. (2008). Exploring the effect of a middle response category on response style in attitude measurement. *Quality & Quantity*, *42*(6), 779–794. <https://doi.org/10.1007/s11135-006-9067-x>.
- Moors, G., Kieruj, N. D., & 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. <https://doi.org/10.1177/0081175013516114>.
- Morren, M., Gelissen, J., & Vermunt, J. (2012). The impact of controlling for extreme responding on measurement equivalence in cross-cultural research. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *8*(4), 159–170. <https://doi.org/10.1027/1614-2241/a000048>.
- Naemi, B. D., Beal, D. J., & Payne, S. C. (2009). Personality predictors of extreme response style. *Journal of Personality*, *77*(1), 261–286. <https://doi.org/10.1111/j.1467-6494.2008.00545.x>.
- Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling*, *14*(4), 535–569. <https://doi.org/10.1080/10705510701575396>.

- O'Muircheartaigh, C., Krosnick, J. A., & Helic, A. (1999). Middle alternatives, acquiescence, and the quality of questionnaire data. Paper presented at the American Association for Public Opinion Research Annual Meeting, St. Petersburg, FL.
- Peer, E., Vosgerau, J., & Acquisti, A. (2014). Reputation as a sufficient condition for data quality on Amazon mechanical Turk. *Behavior Research Methods*, *46*(4), 1023–1031. <https://doi.org/10.3758/s13428-013-0434-y>.
- Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology*, *42*(1), 116–131. <https://doi.org/10.1037/0022-3514.42.1.116>.
- Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *88*(5), 879–903. <https://doi.org/10.1037/0021-9010.88.5.879>.
- Presser, S., & Schuman, H. (1980). The measurement of a middle position in attitude surveys. *The Public Opinion Quarterly*, *44*, 70–85. <https://doi.org/10.1086/268567>.
- Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica*, *104*(1), 1–15. [https://doi.org/10.1016/S0001-6918\(99\)00050-5](https://doi.org/10.1016/S0001-6918(99)00050-5).
- Rammstedt, B., & Krebs, D. (2007). Does response scale format affect the answering of personality scales? Assessing the big five dimensions of personality with different response scales in a dependent sample. *European Journal of Psychological Assessment*, *23*(1), 32–38. <https://doi.org/10.1027/1015-5759.23.1.32>.
- Revilla, M. A., Saris, W. E., & Krosnick, J. A. (2014). Choosing the number of categories in agree-disagree scales. *Sociological Methods & Research*, *43*(1), 73–97. <https://doi.org/10.1177/0049124113509605>.
- Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling and more. Version 0.5–12 (BETA). *Journal of Statistical Software*, *48*(2), 1–36.
- Saris, W. E., & Gallhofer, I. (2007). Design, evaluation, and analysis of questionnaires for survey research. Wiley.
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464 <http://www.jstor.org/stable/2958889>.
- Schwarzer, R., & Jerusalem, M. (1995). *Generalized Self-Efficacy scale*. In J. Weinman, S. Wright, & M. Johnston (Eds.), *measures in health psychology: A user's portfolio. Causal and control beliefs* (pp. 35–37). Windsor: NFER-N.
- Selove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, *52*(3), 333–343. <https://doi.org/10.1007/BF02294360>.
- Shafteel, J., Nash, B. L., & Gillmor, S. C. (2012). Effects of the number of response categories on rating scales. In *Proceedings of the annual conference of the American Educational Research Association* (pp. 1–24).
- Smith Jr., E. V., Ying, Y., & Brown, S. W. (2011). Using the mixed Rasch model to analyze data from the beliefs and attitudes about memory survey. *Journal of Applied Measurement*, *13*(1), 23–40.
- Spector, P. E. (1997). *Job satisfaction: Application, assessment, causes, and consequences*. Thousand Oaks: Sage Publications Inc..
- 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. <https://doi.org/10.1177/0049124112452527>.
- Summerfield, M., Beviit, A., Freidin, S., Hahn, M., La, N., Macalalad, N., et al. (2017). HILDA user manual – Release 16. Melbourne Institute of Applied Economic and Social Research, University of Melbourne.
- Swait, J., & Adamowicz, W. (2001). The influence of task complexity on consumer choice: A latent class model of decision strategy switching. *Journal of Consumer Research*, *28*(1), 135–148. <https://doi.org/10.1086/321952>.
- Tarka, P. (2016). CFA-MTMM model in comparative analysis of 5-, 7-, 9-, and 11-point a/D scales. In A. F. Wilhelm & H. A. Kestler (Eds.), *Analysis of large and complex data* (pp. 553–562). Cham: Springer.
- Tooksoon, H. M. P. (2011). Conceptual framework on the relationship between human resource management practices, job satisfaction, and turnover. *Journal of Economics and Behavioral Studies*, *2*(2), 41–49.
- Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. New York: Cambridge University Press.
- Tourangeau, R., Couper, M. P., & Conrad, F. (2007). Colors, labels, and interpretive heuristics for response scales. *Public Opinion Quarterly*, *71*(1), 91–112. <https://doi.org/10.1093/poq/nf046>.
- Van Vaerenbergh, Y., & Thomas, T. D. (2013). Response styles in survey research: A literature review of antecedents, consequences, and remedies. *International Journal of Public Opinion Research*, *25*(2), 195–217. <https://doi.org/10.1093/ijpor/eds021>.
- Vermunt, J. K., & Magidson, J. (2013). *Technical Guide for Latent GOLD 5.0: Basic, Advanced, and Syntax*. Belmont: Statistical innovations Inc.
- Viswanathan, M., Sudman, S., & Johnson, M. (2004). Maximum versus meaningful discrimination in scale response: Implications for validity of measurement of consumer perceptions about products. *Journal of Business Research*, *57*(2), 108–124. [https://doi.org/10.1016/s0148-2963\(01\)00296-x](https://doi.org/10.1016/s0148-2963(01)00296-x).

- von Davier, M., & Yamamoto, K. (2007). Mixture-distribution and HYBRID Rasch models. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models* (pp. 99–115). New York: Springer.
- Voorpostel, M., Tillmann, R., Lebert, F., Weaver, B., Kuhn, U., Lipps, O., et al. (2010). *Swiss household panel Userguide (1999–2009), wave 11*. Lausanne: FORs.
- Wagner, G. G., Frick, J., & Schupp, J. (2007). The German socio-economic panel study (SOEP). Scope, evolution and enhancements. *Schmollers Jahrbuch*, 127(1), 139–169. <https://doi.org/10.2139/ssrn.1028709>.
- Weathers, D., Sharma, S., & Niedrich, R. W. (2005). The impact of the number of scale points, dispositional factors, and the status quo decision heuristic on scale reliability and response accuracy. *Journal of Business Research*, 58(11), 1516–1524. <https://doi.org/10.1016/j.jbusres.2004.08.002>.
- Weijters, B., Cabooter, E., & Schillewaert, N. (2010). The effect of rating scale format on response styles: The number of response categories and response category labels. *International Journal of Research in Marketing*, 27(3), 236–247. <https://doi.org/10.1016/j.ijresmar.2010.02.004>.
- Weijters, B., Geuens, M., & Schillewaert, N. (2010a). The individual consistency of acquiescence and extreme response style in self-report questionnaires. *Applied Psychological Measurement*, 34(2), 105–121. <https://doi.org/10.1177/0146621609338593>.
- Weijters, B., Geuens, M., & Schillewaert, N. (2010b). The stability of individual response styles. *Psychological Methods*, 15(1), 96–110. <https://doi.org/10.1037/a0018721>.
- Weng, L. J. (2004). Impact of the number of response categories and anchor labels on coefficient alpha and test-retest reliability. *Educational and Psychological Measurement*, 64(6), 956–972. <https://doi.org/10.1177/0013164404268674>.
- Wetzel, E., & Carstensen, C. H. (2014). Reversed thresholds in partial credit models: A reason for collapsing categories? *Assessment*, 21(6), 765–774. <https://doi.org/10.1177/1073191114530775>.
- Wetzel, E., Carstensen, C. H., & Böhnke, J. R. (2013). Consistency of extreme response style and non-extreme response style across traits. *Journal of Research in Personality*, 47(2), 178–189. <https://doi.org/10.1016/j.jrp.2012.10.010>.
- Wetzel, E., Lüdtke, O., Zettler, I., & Böhnke, J. R. (2016). The stability of extreme response style and acquiescence over 8 years. *Assessment*, 23(3), 279–291. <https://doi.org/10.1177/1073191115583714>.
- Winkler, N., Kroh, M., & Spiess, M. (2006). Entwicklung einer deutschen Kurzskaala zur zweidimensionalen Messung von sozialer Erwünschtheit [development of a German short scale for two-dimensional measurement of social desirability]. Discussion paper 579, DIW Berlin. [http://www.diw.de/sixcms/detail.php?id=diw\\_02.c.232162.de](http://www.diw.de/sixcms/detail.php?id=diw_02.c.232162.de)
- Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement*. Chicago: MESA Press.
- Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, 30(3), 187–213. <https://doi.org/10.1111/j.1745-3984.1993.tb00423.x>.
- Zettler, I., Lang, J. W., Hülshager, U. R., & Hilbig, B. E. (2016). Dissociating indifferent, directional, and extreme responding in personality data: Applying the three-process model to self-and observer reports. *Journal of Personality*, 84(4), 461–472. <https://doi.org/10.1111/jopy.12172>.

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## Affiliations

Tanja Kutscher<sup>1,2</sup> · Michael Eid<sup>2</sup>

✉ Tanja Kutscher  
tanja.kutscher@lifbi.de

Michael Eid  
michael.eid@fu-berlin.de

<sup>1</sup> Department of Data Center and Method Development, Leibniz Institute for Educational Trajectories, Wilhelmplatz 3, D-96047 Bamberg, Germany

<sup>2</sup> Department of Psychology, Division of Methods and Evaluation, Freie Universitaet Berlin, Habelschwerdter Allee 45, D-14195 Berlin, Germany