Fachbereich Erziehungswissenschaft und Psychologie

der Freien Universität Berlin

Measuring Job Satisfaction with Rating Scales: Problems and Remedies

Dissertation

zur Erlangung des akademischen Grades

Doktorin der Philosophie

(Dr. phil.)

vorgelegt von

Dipl.-Psych. Tanja Kutscher

Berlin, 2019

Erstgutachter:

Prof. Dr. Michael Eid (Freie Universität Berlin)

Zweitgutachter:

Prof. Dr. Martin Schultze (Goethe-Universität Frankfurt am Main)

Datum der Disputation: 08.07.2019

Acknowledgments

First of all, I would like to thank my advisor Prof. Dr. Michael Eid. Due to his invitation to work with him and be a part of his team, I have become fascinated by applying item response models to improve the quality of measures. Without his invaluable support, inspiration, and constant optimism, I would not have been able to finish this work.

I also would like to thank Prof. Dr. Martin Schultze who kindly agreed to serve as a second advisor for this dissertation. He was also my first contact point for methodological questions.

My special thanks go to Georg Hosoya for helpful discussions about IRT modeling. I would also like to thank Judith Mangelsdorf, Fenne von großen Deters, and Maximilian Bee who supported me in carrying out the experimental online study. I also would like to acknowledge Claudia Crayen who helped me very much with the revision of some of my papers.

Moreover, I also wish to thank all my wonderful colleagues from our work group at Freie Universitaet Berlin for creating a very fantastic atmosphere, making the work more comfortable and productive. It was a pleasure to work with such talented scientists.

And finally, I am very grateful to my friends and my family for their motivation, emotional support, and for doing everything to encourage me to continue and finalize this large project.

Abstract

Job satisfaction is an aspect of cognitive well-being and one of the standard indicators of quality of life. A job satisfaction measure is included in several national panel surveys. The assessment of job satisfaction with a precise and valid measure is a pre-requisite for obtaining accurate analysis results and drawing valid conclusions. However, an inadequately designed response format can impair the way respondents answer the questions, and there is reason to suspect that the 11-point rating scale standardly used in national panel surveys for assessing cognitive well-being could be a problem. Respondents may be overwhelmed by the large number of response categories and, therefore, cope with an increased response burden by using response styles (e.g., overusing particular response categories) and other types of inappropriate category use (e.g., careless responses or ignoring irrelevant or unclear categories). Consequently, data provided by panel surveys may be of reduced quality. Thus, the research in the present dissertation aimed first to investigate whether an 11-point rating scale is adequate for a valid assessment of job satisfaction, one of the relevant life domains. Due to the lack of evidence, the second aim was to examine the performance of mixed polytomous item response theory (IRT) models when applied to detect inappropriate category use under the data condition typical for panel surveys with a job satisfaction measure. The third aim was to study whether a rating scale with fewer response categories may be more optimal to measure job satisfaction. In addition, the fourth aim was to describe the personal profiles of response-style users by means of personality trait, cognitive ability, socio-demographic variables, and contextual factors. It is important to identify these profiles because a person's use of a specific response style can occur consistently across different traits and rating scales and, therefore, is considered a type of disposition.

To examine the adequacy of an 11-point rating scale, we explored patterns of category use in the data on job satisfaction provided by the Household, Income and Labour Dynamics in Australia (HILDA) survey (first wave, n = 7,036). For this purpose, mixed polytomous IRT models were applied. The analyses showed that most respondents (60%) overused extreme response categories (e.g., adopted an extreme response style [ERS]) or the two lowest and two highest categories (e.g., adopted a so-called semi-extreme response style [semi-ERS]), whereas others demonstrated more appropriate response behavior (a so-called differential response style [DRS]). Moreover, all respondents ignored many response categories, especially those who exhibited the ERS and semi-ERS. These findings emphasize the limited adequacy of a long rating scale for assessing job satisfaction due to a large presence of inappropriate category use. Generally speaking, an 11-point rating scale does not allow one to assess fine-grained differences between respondents in their levels of job satisfaction, as intended by the developers of panel surveys. In contrast, this rating scale seems to overburden respondents with superfluous response categories and evoke response styles due to the difficulties they experience by determining the meaning of fine categories. To conclude, a rating scale with fewer response categories may be more optimal.

To address the second aim, a Monte Carlo simulation study was conducted. It included two models: the mixed partial credit model (mPCM; Rost, 1997) and the restricted mixed generalized partial credit model (rmGPCM; GPCM; Muraki, 1997; mGPCM; von Davier & Yamamoto, 2004). These models are suitable for detecting patterns of inappropriate category use. The latter model is more complex and includes freely estimated item discrimination parameters (but which are restricted to be class-invariant). In particular, the simulation study focused on identifying the required sample size for a proper application of these models. In addition, we investigated what information criteria (AIC, BIC, CAIC, AIC3, and SABIC) are effective for model selection. Analysis showed that both models performed appropriately with at least 2,500 observation. By further increasing the sample size, more accurate parameter and standard error estimates could be obtained. Generally, the simulation study revealed that the mPCM performed slightly better than the rmGPCM. Specifically, both models showed estimation problems due to low category frequencies, leading to inaccurate estimates. For the recommended sample size, both the AIC3 and the SABIC were the most suitable. For the large sample sizes (consisting of at least 4,500 cases), both the BIC and CAIC were effective. The AIC, however, was insufficiently accurate.

For the third aim, an experimental study with a between-subject design and randomization was conducted to compare the performance of two short rating scales (with 4 and 6 response categories) with that of a long rating scale (11 response categories) with regard to the presence of inappropriate category use and reliability (*N* = 6,999 employees from the USA). For this purpose, the multidimensional mixed polytomous IRT model was applied. Notably, the results from the simulation study were used at the preparation stage of this study (e.g., regarding the minimum sample size required within an experimental condition). Overall, when the rating scale was short, both the proportion of respondents who used a specific response style and the number of ignored response categories were reduced, indicating less bias in data collected with short rating scales. This finding confirmed the suggestion that some respondents use response styles as an adjustment strategy due to the inadequately large number of response categories offered. Interestingly, the same response styles were present regardless of rating scale length, suggesting that optimizing rating scale length can only partly prevent inappropriate category use. Apparently, a proportion of the respondents use a particular response style due to dispositions.

To attain the fourth aim, the personal profiles of respondents who used a particular response style were investigated with two datasets: (i) a small set of the potential predictors that were available in the HILDA survey (socio-demographic variables and job-related factors); and (ii) several relevant scales and variables (personality traits, cognitive ability, socio-demographic variables, and job-related factors) that were intentionally collected in the experimental study for this purpose. For both datasets, the assignment of respondents to latent classes indicating different response styles was an outcome variable.

The analyses were conducted using multinomial logistic regressions. Therefore, the findings obtained on the basis of the first dataset provided the response-format-specific characteristics of response-style users (for the 11-point rating scale). By contrast, the second analysis allowed to reveal general predictors that explained the use of a particular response style, regardless of rating scale length, whereas response-format-specific predictors explained the occurrence of a response style for a certain rating scale. Specifically, some of the general predictors found for ERS use included a high level of general self-efficacy and self-perceived job autonomy; for non-ERS use, as a tendency to avoid extreme categories, a low need for cognition was the general predictor, indicating that response styles can be caused by dispositions, and therefore they can hardly be prevented by optimizing the features of a rating scale. The predictors specific to a particular response format were then socio-demographic variables, cognitive abilities, and certain job-related factors, suggesting that profiles of respondents who used a particular response style vary depending on the rating scale administrated to collect data. Presumably, these groups of predictors primarily characterize respondents who are inclined to use response styles as an adjustment strategy due to an inadequately designed rating scale.

In sum, an 11-point rating scale was shown to have serious shortcomings, including a high proportion of respondents with response styles and many ignored response categories. Therefore, this rating scale is of limited adequacy for a valid assessment of job satisfaction (and other aspects of cognitive well-being). By contrast, the 4- and 6-point rating scales showed a superior performance with regard to the presence of inappropriate category use. These short rating scales were found to have fewer respondents using response styles and to include almost no redundant response categories. Thus, these shorter rating scales are more adequate for this purpose. Generally, shorter rating scales eliminated the inappropriate category use that is primarily measure-dependent. Nevertheless, the same response styles were present in the data, regardless of rating scale length, suggesting that stable dispositions may be another major cause of response styles. Furthermore, some of these personal characteristics were identified (as general predictors). For example, ERS use could be explained by a high level of general self-efficacy and self-perceived job autonomy. Therefore, any optimizing of the rating scale may not be sufficient to eliminate effects caused by the consistent use of response styles. In this case, statistical approaches of controlling the effects of response styles should be applied. A promising approach for dealing with inappropriate category use are mixed polytomous IRT models.

References

Muraki, E. (1997). A generalized partial credit model. In W. van der Linden & R. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 153-164). New York: Springer.

Rost, J. (1997). Logistic mixture models. In W. van der Linden & R. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 449-463). New York: Springer.

von Davier, M., & Yamamoto, K. (2004). Partially observed mixtures of IRT models: An extension of the generalized partial-credit model. *Applied Psychological Measurement, 28*, 389-406. doi: 10.1177/0146621604268734

Table of Contents

1 INTRODUCTION

Introduction

Several national panel surveys collect representative data from large populations. These data are made accessible to researchers from a range of different research areas. Psychologists can use these data to investigate diverse research topics, such as the intraindividual and interindividual differences in subjective well-being across lifespan, the influence of personality on subjective well-being, or the relationship between income and subjective well-being. To draw accurate and valid conclusions, researchers should ensure that their analyses use unbiased data, however. Thus, the present dissertation addresses the quality of data on cognitive well-being, in particular job satisfaction, that are provided by national panel surveys. One cause of measurement error may be a sub-optimal rating scale (Podsakoff, MacKenzie, Lee, & Podsakoff, 2003), such as can occur if the rating scale consists of many response categories. In the presence of this type of rating scale, respondents can experience difficulties in accurately selecting the most suitable response category. This increased response burden may lead to inappropriate category use (ICU), for example, in terms of misusing certain response categories (response styles [RS]) or ignoring unclear categories.

Today, an 11-point rating scale is the gold standard for measuring cognitive well-being, including job satisfaction (Diener, Inglehart, & Tay, 2013), and it is therefore widely used in national panel surveys. Initially, this response format was used in the German Socio-Economic Panel Study (GSOEP; Richter, Rohrer, Metzing, Nestler, Weinhardt, & Schupp, 2017) by arguing that it allows an effective differentiation among respondents' trait levels, as opposed to shorter rating scales (Cummins & Gullone, 2000). Later, an 11-point rating scale was adopted by other panel surveys, including the Household, Income and Labor Dynamics in Australia survey (HILDA; Summerfield, Bevitt, Freidin, Hahn, La, Macalalad, … & Wooden, 2017) and the Swiss Household Panel study (SHP; Voorpostel, Tillmann, Lebert, Kuhn, Lipps, Ryser, ... & Wernli, 2015). However, the question of whether an 11-point rating scale is able to provide high-quality data has not yet been investigated.

Furthermore, researchers should be aware that even if a rating scale is properly designed, panel data may contain some person-related ICU bias. One explanation may be Krosnick's (1991) concept of satisficing. He differentiates two major types of response behavior: optimizing and satisficing. Optimizing implies that a respondent endeavors to answer accurately and goes through all phases of the response process carefully: He or she correctly interprets the meaning of a question, extensively retrieves all relevant information from memory, integrates that information into a judgment, and thinks carefully about the appropriateness of response categories before endorsing the most reasonable one. Therefore, optimizers generally provide valid responses. In contrast, satisficing occurs when a respondent fails in one of the phases of the response process (e.g., because he or she understands the question superficially, does not figure out the meaning of response categories, or is demotivated or distracted). Therefore, satisficers are inclined to ICU. For instance, some of them overuse particular response categories, regardless of the rating scale design.

The present dissertation had four aims: (1) to determine whether the 11-point rating scale is valid for assessing job satisfaction; (2) to address the performance of mixed item response theory (IRT) models for polytomous data when applied to the data situation typical for national panel surveys, especially regarding measures of cognitive well-being; (3) to identify the optimal rating scale length for improving the psychometric quality of job satisfaction measures included in national panel surveys; and (4) to describe personal profiles of RS users. For the first and third aims, the focus was primarily placed on detecting patterns of ICU. The presence of high ICU may raise concerns about the validity of respondents' responses and therefore suggests that an 11-point rating scale is of limited adequacy for data collection. A well-established approach for detecting patterns of ICU, the mixture polytomous IRT approach, was applied in the studies of the present dissertation to explore ICU in data collected using 11 and fewer response categories.

The introduction next provides a background on the studies of this dissertation. First, it explains the relevance of job satisfaction as an aspect of cognitive well-being and an indicator of quality of life. Second, it describes how cognitive well-being, including job satisfaction, is measured in the context of national panel surveys. Third, the forms of ICU are presented. The fourth and fifth parts summarize previous research on the effects of features of a rating scale, including rating scale length, and the characteristics of respondents with respect to ICU. Sixth, the role of the mixed polytomous IRT models for detecting patterns of ICU is highlighted. Finally, an overview of studies of the present dissertation is given.

1.1 Job Satisfaction As an Aspect of Cognitive Well-Being and an Indicator of Quality of Life

The HILDA survey (Watson & Wooden, 2012) and the GSOEP study (Wagner, Frick, & Schupp, 2007) are some of many national panel surveys that collect annual data on social and economic indicators (e.g., family and household, income and wealth, and the labor market), as well as an increasing number of psychological variables (e.g., the big five personality traits, psychological and subjective well-being, and health). These survey data are made accessible to the broader academic research community and government agencies to support research on diverse multidisciplinary topics and to create a solid basis for evidence-based policy. Thus, the key strength of such survey data is that they allow researchers to measure complex and multifaceted constructs, such as quality of life, using a wide range of objective social indicators (e.g., the gross national income per capita or the level of crime) and subjective indicators (e.g., subjective well-being). Both types of indicators are complementary: objective indicators judge the living conditions of a society, whereas subjective well-being reflects the perspectives of individuals with regard to how they feel and think about their lives (Diener & Suh, 1997). (The book edited by Glatzer, Camfield, Møller, and Rojas [2015] provides a detailed description of the concept of quality of life.) In addition, the cross-national equivalent files (CNEF, based at Cornell University) are available. This data is a result of the cooperation of several research institutes, each responsible for conducting a specific panel survey. This data enables cross-cultural comparisons concerning well-being and its determinants (Frick, Jenkings, Lillard, Lipps, & Wooden, 2007). A comprehensive view of quality of life within a nation and across nations should be useful to guide public policy (Diener et al., 2009). However, an accurate assessment of quality-of-life indicators is required to draw valid conclusions (Diener & Ryan, 2009).

This dissertation primarily addresses measurement issues related to measuring subjective well-being, a key component of quality of life. Subjective well-being is a broad concept that includes cognitive and affective components (Diener, Scollon, & Lucas, 2009). Cognitive well-being refers to individuals' life satisfaction in general and satisfaction with specific life domains (e.g., job satisfaction, satisfaction with social relationships, and marital satisfaction), whereas affective well-being refers to individuals' emotional reactions on their lives (e.g., feeling of happiness or depressed mood) (Diener, Lucas, & Oishi, 2002). Different methodological approaches are required to assess cognitive and affective components of well-being (Eid, 2008). Affective well-being is less stable and depends more on current life events (e.g., Luhmann, Hofmann, Eid, & Lucas, 2012). By contrast, cognitive well-being is relatively stable over time and, therefore, better reflects how individuals enjoy their lives (e.g., Schimmack, Krause, Wagner, & Schupp, 2010). Specifically, this dissertation focuses on cognitive well-being, which is more frequently included in national panel surveys, rather than affective well-being. The measurement of cognitive well-being in panel surveys is usually performed in a uniform manner. Respondents are presented with proposed item sets, each of which measures satisfaction with a particular domain of life. These items sets are offered with the same response format so that methodological issues concerning satisfaction with one life domain can be examined and conclusions can be generalized to other life domains (Diener & Suh, 1997). The present dissertation was primarily limited to one of the major life domains for most adults: job satisfaction. Job satisfaction primarily refers to individuals' cognitive evaluation of specific aspects of their jobs such as working conditions, salary, the work itself, and relationship with supervisors. In addition, it incorporates beliefs about the job and affective experience while on the job (Weiss, 2002). Individuals who enjoy their job are more likely to report a high level of subjective well-being (for a meta-analysis, see Bowling, Eschleman, Wang, 2010; for a review, see Lyubomirsky, King, & Diener, 2005). In an organizational context, the measurement of job satisfaction can be useful in evaluating human resource management, organization of work processes, effectiveness of social responsibility policy, and commitment and the productivity of employees (Judge, Thorensen, Bono, & Patton, 2001). For example, low job satisfaction may be a reason for a high employee turnover, employee absenteeism, psychosomatic illnesses, and lower productivity (e.g., Faragher, Cass, & Cooper, 2005; Wright & Bonnett, 2007). By contrast, employees' high job satisfaction is particularly associated with higher work quality, stronger commitment, and constructive relationships with superiors and colleagues (see Rafferty & Griffin, 2009).

1.2 Measuring Cognitive Well-Being in National Panel Surveys

Among panel surveys, cognitive well-being is measured from the subjective standpoint of respondents (by self-report). An overview of the measures of cognitive well-being in panel surveys is given in Table 1.1. The overall life satisfaction is mostly assessed using a single item: "All things considered, how satisfied are you with life as a whole these days?" Respondents are usually asked to rate this item on a 1-to-10 or 0-to-10 numerical rating scale with verbal labels (*totally dissatisfied* and *totally satisfied*) at the extremes. In addition, respondents' satisfaction with specific life domains is assessed on the same rating scale using single-item or multi-item measures. The typical life domains are financial situation, job, family relationships, home and living environment, and health. The assessment of job satisfaction occurs in panel surveys in an identical way. A single-item such as "All things considered, how satisfied are you with your job overall?" assesses overall job satisfaction. A few additional items measure specific extrinsic aspects (e.g., pay and promotion) and intrinsic aspects of work (e.g., relationship with colleagues and superiors, work itself, and working conditions).

When examining national panel surveys, some specific features of life and job satisfaction measures are remarkable. First, panel surveys rarely include well-established instruments such as the Satisfaction with Life Scale (SWLS; Diener, Emmons, Larsen, & Griffin, 1985), the Michigan Organizational Assessment Questionnaire (MOA; Cammann, Fichman, Jenkins, & Klesh, 1983), or the Job Descriptive Index (JDI; Smith, Kendall, & Hulin, 1969). Clearly, single-item measures and short scales are preferred because of cost, time limit, and low burden for respondents. Nevertheless, the single-item measure of life satisfaction may be considered reliable and valid (Cheung & Lucas, 2014; Lucas & Donnellan, 2012), as it performs very similarly to multi-item scales. The same is also generally applicable to short scales. However, a slightly different construct may be measured with the single-item measure compared to the multi-item ones but they are highly correlated (Cabrita & Perista, 2007). Second, no consistent measure of life or job satisfaction is used across panel surveys. The type of life domains included and the number of items for each of them vary greatly. Furthermore, a variety of rating scales is used for assessing cognitive well-being, ranging from 4-point rating scales (e.g., in the Survey of Health, Ageing and Retirement in Europ [SHARE] or the World Values Survey [WVS]) to 11-point ones (e.g., in the GSOEP study or the HILDA survey). Recent panel surveys have usually adopted an 11-point rating scale without questioning its adequacy because this rating scale has been used in many previous panel surveys. The strong preference for a rating scale with many response categories is mostly based on two arguments: (i) a high number of response categories supposedly allows researchers to more accurately capture fine-grained differences in the rated trait between respondents (Cummins & Gullone, 2000; Voorpostel et al., 2015); and (ii) a higher number of response options is supposed to improve the reliability of short-scales (Preston & Colman, 2000; Scherpenzeel & Saris, 1997).

These ways of measuring cognitive well-being in panel surveys raise methodological concerns, however: the danger of measurement bias due to self-report measures, limited comparability of results across cultures, and ICU due to sub-optimal rating scales. Below, these concerns are discussed in more detail. First, although self-report measures are mostly used in the field of subjective well-being research, researchers need to be aware that every respondent's response cannot be entirely accurate and valid (Diener & Suh, 1997). Empirical evidence suggests that responses to a self-report measure can in some cases be influenced by several factors, such as memory biases, current emotional state, impression management, situational factors, and the ordering of items in the questionnaire (Diener & Ryan, 2009). Therefore, multi-method measurement may be useful to achieve an objective account of cognitive well-being. Second, due to the absence of a common cognitive well-being measure (incl. job satisfaction measure) across panel surveys, variations in satisfaction measures can reflect different concepts between cultures (Diener & Suh, 2000). This absence may impede cultural comparisons of cognitive well-being, such as between Asian and European respondents. In addition, they may understand the meaning of the same response categories differently. Thus, the application of the IRT approach may be useful in comparing responses across cultures (Tay, Diener, Drasgow, & Vermunt, 2011). Finally, as mentioned above, the use of the rating scale may be a concern (Diener et al., 2013). It cannot be ruled out that respondents' responses may reflect interindividual differences in use of the rating scale beyond interindividual differences in the levels of cognitive well-being (Diener et al., 2013). In particular, it may be problematic when a rating scale includes an excessively high number of response categories (e.g., an 11-point rating scale). Some respondents may consider accurately differentiating between many fine-grained categories to be a quite difficult task, increasing the respondent's burden and possibly provoking inappropriate responses (e.g., Tourangeau, Rips, & Rasinski, 2000). Therefore, the presence of ICU in the data can influence the conclusions drawn from analyses (Podsakoff et al. 2003). The literature currently lacks research on the adequacy of a rating scale with many response categories for accurate and valid assessment in panel surveys. The present dissertation focuses on the last measurement issue. As mentioned above, one of its major aims is to determine the adequacy of an 11-point rating scale for the assessment of job satisfaction, primarily using ICU as an evaluation criterion. Another major aim is, if necessary, to identify an optimal design of the rating scale to minimize measure-related ICU.

Table 1.1. Overview of cognitive well-being measures in panel surveys.

Rating scale	**Overall job satisfaction**[1]	**Aspects of job satisfaction** (nr. of aspects)	**Life satisfaction**[1] (scale)	**Domain-specific life satisfaction** (nr. of life domains)
4-point	SHARE ELSA	SHARE (2) WVS (2)	SHARE	SHARE (3)[1] WVS (1)[2]
5-point		LNU (2)	DEAS (SWLS) PSID	
6-point	ECHP[3]	ECHP[3] (6)		ECHP[1] (3)[1]
7-point	BHPS, UKHLS ISSP	BHPS, UKHLS (4)	BHPS, UKHLS ISSP LISS (SWLS) ELSA	BHPS (8)[1] UKHLS (3)[1] ISSP (1)[1]
10-point	EQLS		EQLS WVS	EQLS (5[1], 1[2]) WVS (1)[1]
11-point	GSOEP HILDA LISS MIDUS SHP	HILDA (5) LISS (6) SHP (1)[2]	GSOEP HILDA LISS MIDUS SHP ESS ELSA	GSOEP (7 - 21)[1, 4] HILDA (8[1], 2[2]) MIDUS (4)[1] SHP (2[1], 5[2]) ESS (1[2])

Notes. **SHARE** = Survey of Health, Ageing and Retirement in Europe (Börsch-Supan, 2013), **ELSA** = English Longitudinal Study on Aging (NatCen Social Research, 2012), **WVS** = World Values Survey (Inglehart, Haerpfer, Moreno, Welzel, Kizilova, Diez-Medrano et al., 2014), **LNU** = Swedish Level-of-Living Survey (Swedish Institute for Social Research, 2010), **DEAS** = Deutscher Alterssurvey [German Ageing Survey] (Dittmann-Kohli, Kohli, Künemund, Motel, Steinleitner, & Westerhof, 1997), **PSID** = Panel Study of Income Dynamics (Duffy, Leissou, McGonagle, & Schlegel, 2013), **ECHP** = European Community Household Panel (EuroPanel Users Network, 2004), **BHPS** = British Household Panel Survey (Taylor et al., 2018), **UKHLS** = UK Household Longitudinal Study (Knies, 2017), **ISSP** = International Social Survey Program (Scholz, Jutz, Edlund, Öun, & Braun, 2014), **LISS** = Longitudinal Internet Studies for the Social Science in the Netherlands (Marchand, 2017; Streefkerk, 2017), **EQLS** = European Quality of Life Survey (European Foundation for the Improvement of Living and Working Conditions, 2010), **GSOEP** = German Socio-Economic Panel Study (Richter, Rohrer, Metzing, Nestler, Weinhardt, & Schupp, 2017), **HILDA** = Household, Income and Labour Dynamics in Australia (Summerfield, Bevitt, Freidin, Hahn, La, Macalalad, … & Wooden, 2017), **SHP** = Swiss Household Panel Study (Voorpostel, Tillmann, Lebert, Kuhn, Lipps, Ryser, ... & Wernli, 2015), **MIDUS** = Longitudinal Study of Midlife in the United States (Ryff, Almeida, Ayanian, Carr, Cleary, Coe, & Williams, 2017), **ESS** = European Social Survey (ESS Data Team, 2017). **SWLS** = Satisfaction with Life Scale (Diener, Emmons, Larsen, & Griffin, 1985).

[1] Singe-item measure. [2] Multi-item scale.

[3] Depending on the country, the items are measured with different rating scales. In the ECHP, the data are harmonized to a 6-point rating scale.

[4] The number of life domains measured varies from wave to wave.

1.3 Inappropriate Category Use

This dissertation aims to evaluate the adequacy of rating scale length using the presence of inappropriate category use (ICU)[1] in the data of job satisfaction as a key criterion. The inappropriate category use is "a systematic tendency to respond to a range of questionnaire items on some other basis than the specific item content" (Paulhus, 1991, p. 17). For instance, two respondents with the same latent trait level can select different response categories due to individual RSs: respondent 1 chooses a more extreme response category while respondent 2 selects a less extreme category. Consequently, ICU impedes the valid representation of individuals' latent trait levels by their observed scores in a systematic way (Baumgartner & Steenkamp, 2001; Podsakoff et al., 2003). A growing body of evidence shows that not accounting for the effects of ICU affects the quality of measures. Specifically, the location and shape of response distributions may be biased (see Cheung & Rensvold, 2000; Mõttus, Allik, Realo, Rossier, Zecca, Ah-Kion, ... & Bhowon, 2012; Reynolds & Smith, 2010); researchers can then obtain inflated or deflated correlation and regression coefficients or biased model parameters (see Khorramdel & von Davier, 2014; Moors, 2012; Morren, Gelissen, & Vermunt, 2012; Rossi, Gilula, & Allenby, 2001; Tutz, Schauberger, & Berger, 2018). Furthermore, ICU may destroy the dimensional structure of a trait or attitude (see Aichholzer, 2014; Jin & Wang, 2014) and impair the reliability and validity of scales (see De Jong, Steenkamp, Fox, & Baumgartner, 2008; Dolnicar & Grün, 2009; Jin & Wang, 2014; Weijters, Schillewaert, & Geuens, 2008). Thus, the presence of ICU in the data threatens the validity of conclusions drawn from analyses.

Inappropriate category use[2] is here a generic term that covers different types of inappropriate responses such as RSs, shortcut strategies, social desirability, and careless responses. Table 1.2 provides an overview of these four types of ICU. Below, these types are described in more detail.

(1) Response styles are systematic individual tendencies in response behavior, characterized by choosing certain response categories from a rating scale more frequently than other ones, regardless of the item content. Response styles such as extreme response style (ERS), middle response style (MRS), acquiescence response style (ARS), and disacquiescence response style (DRS) are the most commonly investigated, and thus these are the most often considered when examining the adequacy of a specific rating scale. Responding to the same items, respondents can differ in their response behavior so that some of them use RSs while others use the rating scale as intended (Austin, Deary,

[1] Response bias, RSs, or response sets are also oft-used terms in the literature.

[2] In Chapter 2, the term "inappropriate scale usage" has the same meaning.

& Egan, 2006; Maij-de Meij, Kelderman, & van der Flier, 2008; Meiser & Machunsky, 2008; Wagner-Menghin, 2006; Wu & Huang, 2010). Response styles can also be measure-specific (Cabooter, Weijters, De Beuckelaer, & Davidov, 2017). Depending on the measure applied, different RSs may also exist in the data of the same trait or attitude (for satisfaction with job-related aspects, cf. Carter, Dalal, Lake, Lin, & Zickar, 2011; Eid & Rauber, 2000).

(2) Another form of ICU refers to so-called shortcut strategies (e.g., in terms of ignoring superfluous response categories, preferable use of labeled categories, preferring response categories in a particular area of a rating scale, or exhibiting a narrow response range). Respondents most often use shortcut strategies when they experience difficulties in understanding the meaning of response categories due to an inadequately designed rating scale (Krosnick, 1999). Notably, this type of ICU has been rarely studied, if at all, in the context of the IRT approach (e.g., Andrich, 2010; Wetzel & Carstensen, 2014). In addition, among shortcut strategies, this dissertation is primarily focused on ignoring superfluous response categories, a tendency by which respondents tend to reduce the offered rating scale to a limited number of effective response categories.

(3) Furthermore, the data can contain socially desirable responses. Social desirability refers to the tendency to give responses that are socially approved (Paulhus, 1984). In particular, a high level of socially desirable responses may be expected on questions regarding relevant social issues (e.g., deviant sexual behavior, drug use) or when an assessment occurs in a relevant social situation (e.g., job application process) or when the administration of surveys is conducted in a face-to-face interview (Schwarz, Strack, Hippler, & Bishop, 1991; Zickar & Gibby, 2006). In most panel surveys, questions on cognitive well-being are usually included in the self-completion questionnaire that respondents complete themselves. In this context, a low level of socially desirable responses to job satisfaction can be expected. Therefore, social desirability is concerned beyond the scope of this dissertation.

(4) Finally, approximately 5% of respondents in any sample are inclined to respond to items carelessly (Meade & Craig, 2012). The four most prominent forms of careless responses are inattentive, quick, invariant, and random. Unlike other types of ICU, careless responses should be considered primarily as a source of random measurement error (Baumgartner & Steenkamp, 2001). It is primarily triggered by situational factors (e.g., a lack of motivation to participate in the survey, fatigue, and time pressure). Therefore, the screening of careless responses serves primarily to exclude respondents whose responses are highly inaccurate or invalid (Curran, 2016).

Hence, this dissertation is aimed to evaluate the adequacy of rating scale length with consideration of this multifaceted nature of ICU. For this purpose, the mixture polytomous IRT approach was suitable, which enables different types of ICU to be detected simultaneously (for details, see the section "Methods to Control Inappropriate Category Use").

Table 1.2. Overview of types of inappropriate category use.

Type	Form	Definition or Example
Response styles [1, 2, 3]	Extreme response style	The tendency to choose predominantly the extreme response categories, regardless of content.
	Middle response style	The tendency to prefer use of the middle response category, regardless of content.
	Acquiescence response style	The tendency to agree with items or the preference for positive responses, regardless of content. Also called agreement tendency, yea-saying, or positivity.
	Disacquiescence response style	The tendency to disagree with items or the preference for negative responses, regardless of content. Also called disagreement tendency, nay-saying, or negativity.
	Net acquiescence response style	The tendency to show greater acquiescence than disacquiescence. Also called directional bias.
	Non-extreme response style	The tendency to avoid the extreme response categories or to choose the most non-extreme response categories, regardless of content. Also called the mild response style.
Shortcuts [1, 2, 4; 5, 6, 7]	Ignoring superfluous response categories	For respondents who have a black-and-white-thinking style, the middle category could be superfluous.
	Overuse of labeled categories	For example, with regard to an endpoint labeled rating scale, a respondent preferably uses the extreme categories because they are provided with labeled.
	Response range	The tendency to use a narrow or wide range of response categories around the mean response.
Social desirability [3, 8, 9]	Self-deception	An unconscious tendency to have a distorted perception of reality in an optimistic way to protect self-concept and self-esteem.
	Impression management	A tendency of an individual to deliberately mislead other persons, giving them the most favorable impression.
Careless responding [1, 2, 10, 11]	Inattentive responding	The tendency to select an "incorrect" response category due to insufficient effort to read items carefully.
	Quick responding	The tendency to spend a minimal amount of time when responding to an item.
	Invariant responding	The tendency to select the same response categories for quite a few consecutive items.
	Random responding	The tendency to select response categories on a set of items in an inconsistent way (low within-person reliability index).

Notes. [1] Baumgartner & Steenkamp (2001). [2] Van Vaerenbergh & Thomas (2013). [3] Wetzel, Böhnke, & Brown (2016). [4] Krosnick (1999). [5] Viswanathan, Sudman, & Johnson (2004). [6] Andrich (2010), [7] Wetzel & Carstensen (2014). [8] Paulhus (1984). [9] Zickar & Gibby (2006). [10] Curran (2016). [11] Meade & Craig (2012).

The next point to be considered is how ICU is caused. The answer to this question was expected to reveal ways in which ICU can be eliminated and the quality of data improved. Recent research has emphasized three major causes for ICU:

(1) a deficient measure design (e.g., unclearly formulated items or suboptimal features of a response format), by which respondents may be forced to satisfice rather than to optimize their responses (Krosnick, 1991) due to a sub-optimal response format causing increased response burden (Baumgartner & Steenkamp, 2001; Cox, 1980; Viswanathan, Sudman, & Johnson, 2004);

(2) respondents characteristics (e.g., internal dispositions; Austin et al., 2006; Billiet & Davidon, 2000; Kieruj & Moors, 2013; Krosnick, 1999; Van Vaerenbergh & Thomas, 2013);

(3) contextual factors (e.g., motivation) and cultural norms of response behavior (Baumgartner & Steenkamp, 2001; De Jong et al., 2008; Harzing, Baldueza, Barner-Rasmussen, Barzantny, Canabal, Davila, ... & Liang, 2009; Johnson, Kulesa, Cho, & Shavitt, 2005; Van Herk, Poortinga, & Verhallen, 2004).

Because the empirical part of the present dissertation refers to employees and employers from English-speaking countries (Australia and the USA), the latter cause was beyond the scope of this dissertation. The next two sections discuss how the design of a rating scale and respondent characteristics can affect inappropriate responding.

1.4 Rating Scale As a Cause of Inappropriate Category Use

Rating scales are widely used in social and behavioral research to assess a variety of personality traits and attitudes. This type of response format allows respondents to express their levels of agreement to each scale item by selecting one of the proposed ordered response categories. The popularity of rating scales can be primarily explained by their applicability to diverse types of questions and ease of administration. Usually, rating scales differ in their features (e.g., number of response categories, inclusion or exclusion of a middle category, and verbal and numerical labels of categories) across measures. However, the major criticism of rating scales is their susceptibility to ICU (Cronbach, 1950; Greenleaf, 1992a; Kieruj & Moors, 2010; Morren et al., 2012; Tourangeau & Smith, 1996; Van Vaerenbergh & Thomas, 2013; Weijters et al., 2008). As summarized in Tables 1.3 and 1.4, previous findings show inappropriate responses vary with the features of rating scales. In general, these findings suggest that suboptimal design of the rating scale (e.g., including an excessively large or low number of response categories) is a serious cause for concern.

How the rating scale affects respondents' answers can be understood in consideration of the responding process. According to the response process model proposed by Tourangeau, Rips, and Rasinski (2000), respondents go through four cognitive steps by responding to scale items: understanding

Table 1.3. Summary of previous findings on the effects of rating scale length on inappropriate category use.

Study	3 cat.	4 cat.	5 cat.	6 cat.	7 cat.	8 cat.	9 cat.	10 cat.	11 cat.
Hamby & Levine (2016)		Stronger evidence of the ERS for shorter rating scales							
Kieruj & Moors (2013)			Strong evidence of the ERS Weak evidence of the ARS				Strong evidence of ERS Weak evidence of ARS		
Kieruj & Moors (2010)			No MRS for shorter rating scales Evidence of the ERS				Increase of MRS for longer rating scales Evidence of ERS		
Moors (2008)			Evidence of the ERS						
Weijters, Cabooter, & Schillewaert (2010)		Higher level of the ERS for shorter rating scales		Higher level of MR for longer rating scales No effect on NARS for longer rating scales					
Harzing et al. (2009)			Higher level of the ERS and MRS		Higher level of the ARS and DRS				
Hui & Triandis (1989)			Higher level of the ERS (for Hispanics)						
Clarke (2000a; 2000b)	Strong decrease of the ERS			Slight reduction of the ERS					
Wakita, Ueshima, & Noguchi (2012)					Evidence of ignored categories[1]				
Weathers, Sharma, & Niedrich (2005)						Evidence of invariant responses			

Notes. ERS = extreme response style; MRS = middle response style; ARS = acquiescence response style; DRS = discquiescence response style; MR = misresponse to reversed items; NARS = net acquiescence response style.

[1] Several application studies of the mixture IRT approach also reported the presence of ignored categories for 6- and 7-point rating scales (e.g., Eid & Rauber, 2000).

The cells marked in gray represent different rating scale lengths investigated in a particular study.

Table 1.4. Summary of previous findings on the effects of other features of the rating scale on inappropriate category use.

	Labeling	
Study	**Fully labeling**	**Endpoint labeling**
Moors, Kieruj, & Vermunt (2014)	Strong evidence of the ERS No effect on the ARS	Strong evidence for the ERS; higher level of the ERS No effect on the ARS
Weijters, Cabooter, & Schillewaert (2010)	Higher level of the NARS	Higher level of the ERS and MR
Lau (2007)	No effect on the ERS	No effect on the ERS

	Numbering		
	Increasing positive values	**Negative and positive values**	**No numbering**
Cabooter, Weijters, Geuens, & Vermeir (2016)	Higher level of the ERS	Higher level of the ARS	
Moors, Kieruj, & Vermunt (2014)	Strong evidence of the ERS No effect on the ARS	Strong evidence of the ERS; higher level of the ERS No effect on the ARS	Strong evidence of the ERS No effect on the ARS
Schwarz Knäuper, Hippler, Noelle-Neumann, & Clark (1991)	Higher level of the ERS		

	Inclusion of the middle category	
	Even-numbered rating scales	**Odd-numbered rating scales**
Weijters, Cabooter, & Schillewaert (2010)	Higher level of the ERS and MR	Higher level of the NARS
O'Muircheartaigh, Krosnick, & Helic (1999)	No effect on the ARS	No effect on the ARS No effect on the use of shortcut strategies

Notes. ERS = extreme response style; MRS = middle response style; ARS = acquiescence response style; DRS = discquiescence response style; MR = misresponse to reversed items.

The cells marked in gray represent different types of numbering of response categories in a particular study.

the question, retrieving all associated information from the memory, aggregating this information into a judgment, and reporting the judgment by selecting the matching category from the offered response format. For most respondents of at least average intelligence level without any cognitive disabilities, the first three steps should not provoke any inappropriate responding, provided that items are formulated in clear, simple, and accessible language. However, the last step may be a potential source of an ICU due to a sub-optimally designed rating scale, such as one that is not adjusted to the respondents' thinking complexity and discrimination ability (Arce-Ferrer, 2006; Baumgartner & Steenkamp, 2001; Cox, 1980; Viswanathan et al., 2004). Thinking complexity (e.g., black-and-white-thinking or sophisticated thinking) refers to the number of subjectively meaningful categories a respondent draws upon in thinking about the content of an item (Naemi, Beal, & Payne, 2009; Viswanathan et al., 2004). Discrimination ability refers to the ability of the respondent to accurately decode the meaning of the response categories of the rating scale offered. This ability is also relevant for successfully matching subjective categories and offered ones (Baumgartner & Steenkamp, 2001; Cox, 1980; Krosnick, 1991). If respondents fail to match these categories, they may experience cognitive overload, leading to satisficing in the form of ICU (Greenleaf, 1992a; Hamby & Levine, 2016; Krosnick, 1991; Swait & Adamowicz, 2001; Viswanathan et al., 2004; Weathers, Sharma, & Niedrich, 2005). As described above, panel surveys often assess cognitive well-being using rating scales with a high number of response categories and endpoint labeling. In this scenario, respondents may experience difficulties in determining the meaning of unlabeled response categories located between two labeled extremes, interpret them differently, or consider some of them redundant. If a long rating scale with endpoint labeling produces a large amount of ICU, its use in panel surveys would raise doubts concerning the validity of conclusions drawn from analyses (Billiet & McClendon, 2000; DeVellis, 2016; Revilla, Saris, & Krosnick, 2014; Weathers et al., 2005). Hence, the validation of this rating scale intended in the present dissertation may clarify whether or to what extent this rating scale is appropriate for collecting high-quality data.

The oldest technique for preventing a measure-related ICU is to design an optimal rating scale for a particular measure at the development or validation stages (Cox, 1980; Greenleaf, 1992a; Krosnick, 1999; Weijters et al., 2008).[3] The primary requirement for an optimal rating scale is its capacity to transfer the full information that respondents possess about a specific trait or attitude without overwhelming their cognitive ability (Cox, 1980; Krosnick, 1999; Lozano, García-Cueto, & Muñiz, 2008; Viswanathan et al., 2004). Thereby, both high reliability and validity and the absence of bias, or the presence of only a negligible amount of bias, due to ICU are predominantly considered as the key evaluation criteria (Cox, 1980; Kieruj & Moors, 2010; Lee & Paek, 2014; Podsakoff et al., 2003; Weathers et al., 2005). With regard

[3] Item wording as a further source of ICU bias (for details, see Podsakoff et al., 2003) is beyond the scope of this dissertation because in the panel surveys the items of cognitive well-being (incl. job satisfaction) are written in plain language.

to ICU, previous studies show that shorter rating scales (but not excessively short) could be more suitable for panel studies than are long rating scales. This suitability can be confirmed with the following evidence. First, as reported by empirical studies that applied IRT models to examine the psychometric properties of a particular measure, short rating scales generally outperform long rating scales in terms of, for example, sufficient coverage of the latent trait continuum and the correct order of equidistant categories (e.g., Freund, Tietjens, and Strauss, 2013; Khadka, Gothwal, McAlinden, Lamoureux, & Pesudovs, 2012). In contrast to short rating scales, long rating scales do not adequately reflect the continuity of a latent trait variable if some unordered categories occur (Meiser & Machunsky, 2008). Second, further studies examining the effect of rating scale length on reliability and validity warn against the use a rating scale with less than four and more than six or seven response categories (e.g., Culpepper, 2013; Lee & Paek, 2014; Lozano et al., 2008; Maydeu-Olivares, Kramp, García-Forero, Gallardo-Pujol, & Coffman, 2009). Excessively short rating scales produce low reliability and validity values, whereas excessively long rating scales show artificially high reliability and validity values when ICU effects are not eliminated (Chang, 1994; Jin & Wang, 2014; Revilla et al., 2014; Tarka, 2016). Finally, most respondents can accurately differentiate between no more than six or seven units (Miller, 1956). Thus, an excessively high number of response categories may increase the risk of respondents to choose a less suitable category.

However, previous empirical evidence has not provided any universal rating scale for all measurement situations. The features of an optimal rating scale can vary across constructs, target populations, and researchers' intentions (Cox, 1980; Dolnicar & Grün, 2009; Khadka et al., 2012; Kieruj & Moors, 2010; Weijters, Cabooter, & Schillewaert, 2010). For this reason, the present dissertation focuses on only one trait, namely job satisfaction. Based on the evidence presented above, we expect shortening rating scale length when assessing job satisfaction in panel surveys to reduce the extent of inappropriate responding. However, optimizing rating scale length cannot prevent ICU that is person-dependent and consistently occurs regardless of the rating scale design, as shown in Tables 1.3 and 1.4 (Kieruj & Moors, 2010; 2013; Moors, 2008; Moors, Kieruj, & Vermunt, 2014). Thus, this dissertation aims to identify an optimal rating scale length for assessing job satisfaction, an aspect of cognitive well-being, and to describe the profiles of respondents who use a particular RS consistently.

1.5 Personal Profiles of Inappropriate Category Users

Previous research has provided evidence that ICU may be considered a type of disposition, suggesting that some respondents are prone to use certain response strategies regardless of either item content or the response format (cf. Böckenholt, 2012; Kieruj & Moors, 2013; Moors et al., 2014; Weijters, Geuens, & Schillewaert, 2010a; 2010b). For instance, empirical studies that applied the mixture IRT approach revealed that within each sample, a proportion of the respondents use the response format inappropriately (e.g., Austin et al., 2006, Carter et al., 2011; Eid & Rauber, 2000; Eid & Zickar, 2007; Maij-de Meij et al., 2008; Meiser & Machunsky, 2008; Wagner-Menghin, 2006; Wu & Huang, 2010). Furthermore, several experimental studies have reported that ICU (e.g., in the form of the ERS) is present in the data, regardless of the features of a rating scale, such as rating scale length (e.g., Clarke, 2000a; 2000b; Harumi, 2011; Kieruj & Moors, 2010; 2013; Moors, 2008), labels for response options (e.g., Harumi, 2011; Lau, 2007; Moors et al., 2014), and (ascending, descending or absent) numerical labels (e.g., Liu & Keusch, 2017), as shown in Tables 1.3 and 1.4. Moreover, the use of RSs is largely consistent across personality and attitude scales within large-assessment surveys, specifically when most of them were administered with the same rating scales (see Danner, Aichholzer, & Rammstedt, 2015; Weijters et al., 2010a; Wetzel, Carstensen, & Böhnke, 2013; Zettler, Lang, Hülsheger, & Hilbig, 2015). Relatively stable use of RSs is also found in the longitudinal data when participants answered the same questions at multiple points of time over a period varying between of one month to eight years, depending on the study (see Billiet & Davidov, 2008; Danner et al., 2015; Javaras & Ripley, 2007; Keller & Koller, 2015; Kieruj, 2012; Weijters et al., 2010b; Wetzel, Lüdtke, Zettler, & Böhnke, 2016). Thus, this evidence suggests that in some cases, ICU is respondent-dependent and may be traced back to dispositions.

So far, little is known about the determinants of person-related ICU. Previous research on this issue investigated how different predictors such as personality traits and attitudes, cognitive ability, and socio-demographic variables contribute to explain person-related ICU but yielded inconclusive findings (see Table 1.5).[4] These studies were mostly on ERS use. Despite ERS being the most-investigated RS, there is mixed evidence concerning personality traits. For instance, some studies have reported that respondents preferring extreme categories are high in neuroticism (Baumgartner & Steenkamp, 2001; Hernández, Drasgow, & González-Romá, 2004), whereas other studies showed exactly the opposite (Gerber-Braun, 2010). In addition, inconsistent findings concern the role of cognitive ability for the person-related ERS use. Some studies have rejected the fact that respondents with a low level of cognitive ability are especially inclined to consistent and stable ERS use (e.g., Naemi et al., 2009). Furthermore, an inconsistent picture arises regarding the effect of socio-demographic variables on the ERS. Although it would be wrong to say that respondents who are women, older, and have a low education level primarily prefer extreme categories (for no effect of education, see Kieruj & Moors, 2013; Moors, 2008; for no gender differences, see Clarke, 2000a; 2000b; Greenleaf, 1992b; Kieruj & Moors, 2013; Moors, 2008; Naemi et al., 2009; for ERS use by young people, see Austin et al., 2006; Gerber-Braun, 2010), socio-

[4] Table 1.5 also provides an overview of the effect of contextual factors (e.g., job-related variables, motivation, task difficulty, and external distractions). Job-related variables are included because the job context can affect how respondents use the response format by answering items about job satisfaction. Further situational variables are considered because of their impact on the use of shortcut strategies by responding.

demographic variables notably explain up to 8% of the variance in RSs (Meisenberg & Williams, 2008; Weijters et al., 2010b). By contrast, the boosting effect of cognition-related traits and attitudes such as intolerance to ambiguity or simplistic thinking on ERS use is not to be doubted. These effects account for nearly a quarter of the variance in extreme responses after controlling for socio-demographic characteristics and cognitive levels (Naemi et al., 2009).

Regarding other traditional RSs, only rudimentary but less inconsistent findings are available. For instance, MRS users are more introverted, more neurotic, less conscientious, and inclined to use tactics of impression management (Gerber-Braun, 2010; Hernández et al., 2004; Kulas & Stachowski, 2013). Moreover, MRS users are mostly low-educated individuals, suggesting an association between MRS and limited cognitive ability (Böckenholt, 2012; Weijters et al., 2010b). The use of this response style becomes more common in older people (Weijters et al., 2010b), but does not differ by gender (Khorramdel & von Davier, 2014). With regard to the ARS and DRS, these RSs are more commonly found in optimistic respondents (Pedersen, 1967). Similar to other RSs, the ARS and DRS are negatively related to cognitive ability and intelligence (Zhou & McClendon, 1999) and rigid mental organization (Knowles & Nathan, 1997). Conflicting results have been found only in research on the relationship between ARS or DRS and socio-demographic factors. Some studies have suggested that the ARS or DRS was used more often by respondents who were women, older, and had a low level of education (e.g., Billiet & McClendon, 2000; Carter et al., 2011; Krosnick, 1991; Meisenberg & Williams, 2008; O'Muircheartaigh, Krosnick, & Helic, 1999; Weijters et al., 2010a; 2010b). However, other studies could not confirm these findings (see Moors et al., 2014).

The use of shortcut strategies has been rarely investigated (Table 1.5). The available evidence highlights the relevance of cognitive levels, cognition-related traits and attitudes (e.g., need for cognition), and situational factors for their use. This evidence indicates that cognitive overload or indifference to the context of the questionnaire are the primary reasons for the use of this type of ICU (e.g., Cacioppo & Petty, 1984; Krosnick, 1991).

The remarkable inconsistency of the presented findings on the determinants of the ICU may in part be attributed to different statistical methods employed to measure bias caused by ICU (for an overview of statistical approaches, see the section "Methods to Control Inappropriate Category Use"). In addition, the set of included predictors varied strongly across studies. In most cases, a few predictors were used to explain a particular RS (e.g., exclusively socio-demographic variables or exclusively cognitive ability). The next possible cause of the inconsistency in the literature could be that different response formats and constructs were used in studies to measure RSs. As reported above, both features of a rating scale and constructs can evoke different types of ICU. Thus, to overcome these shortcomings, a large set

Table 1.5. Summary of previous findings on the effects of respondent characteristics and situational factors for inappropriate category use.

	Response style			
Predictors	**ERS**	**MRS**	**ARS or DRS**	**Shortcut strategies**
Cognitive ability	No effect [1, 2, 3]; ↓ [4, 5, 6, 7] Verbal cognitive ability and intelligence: ↓ [8] Cognitive complexity: ↓ [24]		Verbal cognitive ability, intelligence, and criticalness: ↓ [9, 10, 11, 12, 13, 14] Cognitive complexity: ↓ [15, 24]	Discrimination ability: ↓ [16] Cognitive complexity: ↓ [17, 18]
Personality traits	Extraversion: ↑ [8, 19, 20, 21] Conscientiousness: ↓ [22, 23]; ↑ [8, 19] Agreeableness: — [19, 21] Openness: ↓ [21]; — [19, 21] Neuroticism: ↑ [24, 25, 26]; ↓ [8]	Extraversion: ↓ [8, 25, 27] Conscientiousness: ↓ [8] Neuroticism: ↑ [8, 25, 27] Impression management: ↑ [25]	Optimism: ↑ [28] Ability to cooperate: ↓ [28] Extraversion: — [20] Viewing themselves as intelligent: — [20] Neuroticism: ↑ [24]	
Cognition-related traits and attitudes	Intolerance of ambiguity: ↑ [2, 24] Simplistic thinking: ↑ [2] Certainty about own attitude: ↑ [24] Decisiveness: ↑ [2] Rigidity and dogmatism: ↑ [24]	Ambivalence and indifference: ↑ [29] Certainty about own attitude: ↓ [24]		Need for cognition: ↓ [30, 31]
Socio-demographic variables	Age: children and older people [5, 24, 32, 33] older people [20, 21, 34, 35, 36, 37] young people [8, 19, 38] Gender: females [17, 19, 32, 37, 39]; males [21] — [2, 4, 6, 20, 35, 38, 40, 41, 42, 43, 44] Education: low level [17, 35, 37, 44, 45]; — [20, 38] Socio-economic status: ↓ [35, 54, 55]	Age: older people [37] Gender: — [43] Education: low level [34, 37]	Age: older people [46, 37, 47, 45, 48, 49] — [36] Gender: females [37, 49] — [36] Educational level: low level [45, 47, 48, 49, 50, 51, 52, 53] — [36] Socio-economic status: ↓ [54, 55]; — [56]	

	Response style			
Predictors	**ERS**	**MRS**	**ARS or DRS**	**Shortcut strategies**
Job-related factors	Job position: low leadership level [17, 57] Job tenure: ↑ [17]	Self-reported level of competence: ↓ [29]		
Situational characteristics	Importance of the topic: ↑ [24, 58, 59]	Importance of the topic: ↓ [24]	Importance of the topic: − [20] Motivation or fatigue: ↓ [47, 60] Time pressure: ↑ [24] External distractions: ↑ [24]	Lack of experience and knowledge about the topic: ↑ [47] Task difficulty: ↑ [47] Time pressure: ↑ [47] Motivation: ↓ [17] External distractors: ↑ [47]

Notes. ↑ stands for a positive effect of a particular correlate on the use of a specific RS. ↓ stands for a negative effect of a particular correlate on the use of a specific RS. − stands for no effect of a particular correlate on the use of a specific RS.

[1] Kerrick (1954); [2] Naemi, Beal, & Payne (2009); [3] Zuckerman & North (1961); [4] Brengelmann (1960); [5] Das & Dutta (1969); [6] Light, Zax, & Gardiner (1965); [7] Wilkinson (1970); [8] Gerber-Braun (2010); [9] Elliott (1961); [10] Forehand (1962); [11] Gudjonsson (1986); [12] Gudjonsson (1990); [13] Gudjonsson & Clare (1995); [14] Zhou & McClendon (1999); [15] Knowles & Nathan (1997); [16] Miller (1956); [17] Eid & Rauber (2000); [18] Krosnick & Alwin (1987); [19] Austin, Deary, & Egan (2006); [20] Kieruj & Moors (2013); [21] Meiser & Machunsky (2008); [22] Ashton & Lee (2007); [23] Zettler, Lang, Hülsheger, & Hilbig (2015); [24] Baumgartner & Steenkamp (2001); [25] Hernández, Drasgow, & González-Romá (2004); [26] Lewis & Taylor (1955); [27] Kulas & Stachowski (2013); [28] Pedersen (1967); [29] Dubois & Burns (1975); [30] Cacioppo & Petty (1982); [31] Cacioppo & Petty (1984); [32] De Jong, Steenkamp, Fox, & Baumgartner (2008); [33] Hamilton (1968); [34] Böckenholt (2012); [35] Greenleaf (1992b); [36] Moors, Kieruj, & Vermunt (2014); [37] Weijters, Geuens, & Schillewaert (2010b); [38] Moors (2008); [39] Berg & Collier (1953); [40] Bachman & O'Malley (1984); [41] Clarke (2000a); [42] Clarke (2000b); [43] Khorramdel & von Davier (2014); [44] Marin, Gamba, & Marin, 1992; [45] Meisenberg & Williams (2008); [46] Carter, Dalal, Lake, Lin, & Zickar (2011); [47] Krosnick (1991); [48] Mirowsky & Ross (1991); [49] O'Muircheartaigh, Krosnick, & Helic (1999); [50] Billiet & McClendon (2000); [51] McClendon (1991); [52] Schuman & Presser (1981); [53] Weijters, Geuens, & Schillewaert (2010a); [54] Carr (1971); [55] Ross & Mirowsky (1984); [56] Calsyn, Roades, & Calsyn (1992); [57] Ross, Steward, & Sinacore (1995); [58] Warr & Coffman (1970); [59] Gibbons, Zellner, & Rudek, 1999); [60] Krosnick & Presser (2010).

of determinants of ICU should be investigated in a systematic way, as will be implemented in the present dissertation (Chapter 4). Researchers should be aware that person-related ICU is independent of measures, and, therefore, cannot be prevented by optimizing the rating scale. Instead, its effect can be eliminated by applying a suitable statistical approach (e.g., the mixture IRT approach).

1.6 Methods to Control the Effects of Inappropriate Category Use

Because RSs can seriously impair the quality of trait or attitude measures and because optimizing the measure alone probably cannot completely prevent it from occurring, a solution is required to eliminate RS effects from the data. Table 1.6 presents several statistical approaches provided in the research literature to measure RSs and control for their effects. In general, these approaches can be classified in several different ways (for references, see Table 1.6):

- with regard to manifest or latent method to measure RSs,
    - *non-model-based approaches* (e.g., calculating a sum-score index for a specific RS on the basis of scale items or a separate set of heterogeneous items) and
    - *model-based latent variable approaches* (e.g., confirmatory factor analysis with a latent method factor, latent class factor analyses, multi-process IRT models, multidimensional IRT models, mixture IRT models, and extensions of IRT models by a random-effect factor);
- with regard to type of scaling of the latent RS variable,
    - approaches that assume *RS to be a categorical variable* (e.g., latent class factor models, or mixed IRT models) and
    - approaches that assume *RS to be a continuous variable* (e.g., multidimensional IRT models);
- with regard to the specificity of RS to be measured,
    - approaches that measure *a scale-specific RS* by using the same items to measure both the substantive trait and the RS (e.g., diverse IRT models) and
    - approaches that measure *a generalizable RS* using an additional set of items (e.g., approaches based on RS indices);
- with regard to type of scaling of items used to measure RSs,
    - approaches that require *interval data* to measure and control RSs (e.g., approaches based on RS indices) and
    - approaches that can model *categorical data* (e.g., latent class factor models, or diverse IRT models);
- with regard to the number of RSs to be modeled,
    - approaches that measure *a specific RS* (e.g., confirmatory factor analysis with a latent method factor, or IRT models with a random-effect factor) and

- approaches that can measure *different RSs* simultaneously (e.g., approaches based on RS indices, multidimensional IRT models, or mixed IRT models).

For further classifications of statistical methods to assess and control ICU, see Van Vaerenbergh and Thomas (2013) and Wetzel, Böhnke, and Brown (2016).

In addition, most of these approaches are so-called *ad hoc* methods, meaning that their application requires an a priori assumption about what RSs are present in the data and a clear definition of these RSs to model them within a particular statistical approach. For instance, if a researcher assumes that some respondents are inclined to the ERS, he or she can operationalize this RS in one of two ways: (i) as a respondent's preference to overuse the lowest and the highest response categories only or (ii) as a respondent's tendency to overuse two lowest and two highest response categories (Jin & Wang, 2014). The first operationalization conforms to the traditional definition of the ERS.[5] Given data collected using a rating scale with many response categories, the question may arise concerning which operationalization is suitable for specifying the RS factor. However, the primary shortcoming of *ad hoc* approaches is their low usefulness for detecting RSs in an exploratory way. This can have two serious consequences:

(1) There is no guarantee that all RSs existing in the data can be eliminated by applying a particular *ad hoc* method (Morren et al., 2012).
(2) An operationalization of RSs within an *ad hoc* model (e.g., by using a scoring function or RS-indices) can influence how accurately the latent person parameters (in the IRT framework) or latent trait values of a substantial trait (in the SEM framework) are adjusted for RS effects (Jin & Wang, 2014). For example, the adjustment of latent person parameters or latent trait values may be limited if a researcher specifies ERS factor according to the traditional definition when respondents' responses contain ERS effects, as described in the second definition above.

In the present dissertation, mixed polytomous IRT models were applied. A distinguishing advantage of these models is that they overcome the limitation of *ad hoc* approaches and detect patterns of an ICU a posteriori. Therefore, the mixture IRT approach can be classified as one of the so-called *post hoc* methods. In this case, researchers are not required to have specific hypotheses about what types of ICU are present in the data. When a mixed IRT model is applied to items that measure a trait or an attitude of interest, latent classes can be identified that differ systematically in their category use (Rost, 1997). Both RSs that occur and the actual number of response categories used by respondents within

[5] The same refers to the MRS. Traditionally, MRS has been defined as the tendency to overuse the middle response category, regardless of content. For long rating scales, both the overuse of the middle category and adjacent categories around the middle category can be considered to be MRS.

Table 1.6. Overview of statistical approaches for controlling inappropriate category use

Statistical approach	Description and representative models	Advantages and disadvantages
Approaches based on RS indices	These approaches require that RS indices are built in the first step. The most common techniques for calculating RS indices are listed below. In the next step, RS indices can be included in a model in the form of predictors, covariates, or indicators of a latent RS factor. Techniques to build RS indices: - **Count procedure** is the easiest technique to build sum score RS indices by counting the number of agreements, disagreements, extreme responses, and midpoint responses (1, 2). - **Representative indicators for response style (RIRS) technique** is based on the count procedure and requires a sufficient set of lowly intercorrelated, maximally heterogeneous items, so-called *representative indicators for RSs* (up to 10–14 items), to obtain reliable and valid RS indices (3, 4). - **Greenleaf's "contentless" RS measure** allows one to build RS indices based on *any* set of uncorrelated items (5). Examples of approaches based on RS indices: - Linear regression (3, 5, 6) - ANCOVA (7) - CFA with RS factors or multidimensional IRT models including RS factors (4, 8, 9, 10, 11) - Multilevel CFA with RS-factors or multi-group CFA with RS factors to be applied, for example, to cross-cultural comparisons (3, 4, 10, 12)	(+) Indices for diverse RSs can be calculated (e.g., ERS, MRS, ARS, DARS, NARS). (+) The intensity of RSs can be measured. (+) RS indices can be used as covariates in subsequent analyses for the validity check. (+) Depending on a correction model, manifest sum scores or latent trait values are adjusted for the effects of considered RSs (with the exception of RS indices measured by the same scale items as used for a measure of a trait or attitude of interest; in this case, RS variance and trait variance are confounded). (–) An additional set of heterogeneous (low-correlated) items is required (unless the same items are used to measure RSs and the substantive trait). (–) The reliability and validity of RS indices depend on the psychometric quality of representative indices for RSs. (–) Only a priori-defined RSs can be measured and controlled. (–) RS indices allow measurement of only the general RS effect (unless the same items are used to measure RSs and the substantive trait; in the latter case, scale-specific RS effects are measured). (–) RS indices provide equal weight to all items, and, therefore, they do not reflect item-specific RS effects. (–) The most correction methods (e.g., linear regression) assume a linear relationship between the trait and the RSs.
CFA with a latent method factor	The measurement model includes an additional method factor that allows for measuring ARS separately from the trait factor (13).	(+) No additional items are necessary.

Statistical approach	Description and representative models	Advantages and disadvantages
		(–) Scale-specific ARS can be measured. (–) It works only with balanced scales (with include positively and negatively worded items). (–) Respondents with high trait levels cannot be separated from respondents with high ARS levels.
Unidimensional IRT models with a random-effect factor	**Extensions of polytomous IRT models by including additional random threshold parameters** that reflect individual category use in the form of shrinkage or expansion of the rating scale and account for RSs. Examples: - The GRM with random thresholds (14) - The proportional threshold model (15) - The random-effects rating scale model (16) - The random-effect generalized rating scale model (17) - The PCM with a random ERS factor (18) **Extension of the factor analysis model by including an additional multiplicative person parameter** that allows to model the individual category use. Parametrization as an FA model and an IRT model is possible. Example: - The factor-analytic model with an additional random person parameter (19).	(+) No additional items are necessary. (+) Calculating ERS intensity occurs from distances between adjacent thresholds. Small distances indicate an individual tendency to ERS use. (+) Individual response process can be depicted graphically. (–) Only accounting for scale-specific ERS is possible (in some models, also for the MRS). (–) Respondents with high trait levels cannot be separated from respondents with high ERS levels.
Multidimensional IRT models	Multidimensional IRT models allow for the simultaneous measurement of multiple substantive and RS traits as different dimensions. When defined as compensatory models, in which latent trait and latent RS traits are measured with the same items, substantive traits and RS traits additively contribute the item response. Within these models, RSs can be variously defined:	(+) No additional data is necessary. (+) Models are applicable to categorical data; RSs are assumed as continuous latent random variables. (+) When applied with anchoring vignettes, any RS type and multiple RSs (e.g., ERS, MRS, ARS, DRS, and SDR) can be modeled simultaneously. (+) When applied with anchoring vignettes, separate effects of the RS trait and the substantive trait can be modeled.

Statistical approach	**Description and representative models**	**Advantages and disadvantages**
	- The same set of items is used to measure both the substantive trait and the RS; thereby, items are recorded into RS indicators in accordance with the definition of a particular RS. - For each item, anchoring vignettes that represent expected heterogeneity of individuals' category use (in other words, different RSs) are written. An anchoring vignette contains specified category slope parameters for a specific RS trait. For example, when ERS trait is included in the multidimensional IRT model and items are measured using a 4-point rating scale, the anchoring vignette may be as follows: [1.5, -1, -1, 1.5]. The data from anchoring vignettes are then incorporated into the multidimensional IRT model as indicators of RSs. Examples: - The multidimensional NRM of the ERS (20, 21) or that of multiple RSs (22, 23) - The multidimensional PCM of multiple RSs (11) - The multidimensional unfolding model of multiple RSs (24) - A multi-group extension of the multidimensional IRT models of multiple RSs to be applied, for example, for cross-cultural comparisons (25, 26)	(+) Modells allow flexibility in specifying RSs: (i) as scale-specific RSs or RSs that are generalizable across scales when more than one substantive trait is measured; (ii) as item-specific RS effects or an equal RS effect of all items; or (iii) as RSs which are uncorrelated or correlated with traits. (–) Only a priori-defined RSs can be measured and controlled.
Multi-process IRT models (IR-tree models)	These models assume a stage-wise response process consisting of different hierarchical sub-processes in which respondents engage when answering items. The number of sub-processes depends on the number of response categories. When applied with a long rating scale, four sub-processes can be generally modeled: 1. Indifference (whether a respondent prefers to use a middle category) 2. Direction (referring to agreement or disagreement with an item) 3. Intensity of agreement or disagreement (whether a respondent prefers to select an extreme category or one of less extreme categories) 4. Central tendency (referring to the selection of one of categories in the middle area of the rating scale).	(+) No additional data is necessary. (+) Applicable to categorical data; RSs are assumed as continuous latent random variables. (+) Detecting heterogeneous scale usage is possible. (+) Both person-specific and item content-specific response processes can be identified. (–) Only the ERS and MRS can be modeled. (–) The application of these models requires that researchers have assumptions about the underlying response process respondents proceed by mapping their judgments to the rating scale.

Statistical approach	Description and representative models	Advantages and disadvantages
	The modeling of a particular sub-process occurs with a set of binary pseudo-items that are recorded from the scale items. The application of the multi-process model provides for each respondent person parameter estimates of all sub-processes considered in the model. Furthermore, the multi-process models allow distinguishing between RS effects that are person-related and RS effects that are content-related. Examples: - The two-decision model, suitable for the even-numbered rating scales (27) - The multi-process GRM (28) - The four-process models, suitable for the rating scale with many response categories (29) - The tree-based IR models (30, 31)	
Latent class factor models (LCFM)	As an extension of the latent class approach by an additional latent RS factor. Latent classes can be identified with regard to the substantive trait and the RS. Examples: - Confirmatory LCFM (32, 33, 34, 35, 36) - Latent class bilinear multinomial logit model (37)	(+) No additional items are necessary. (+) Models are applicable to categorical data. (+) RSs can be a priori-defined by weighting loadings or post hoc interpreted from estimated loadings. (–) Only scale-specific ERS and ARS can be modeled. (–) RS is assumed as the categorical variable.
Mixed IRT models	This group of models allows detecting latent classes which systematically differ in their response patterns. Examples: - The mixed rating scale model (38) - The mixed PCM (38, 39, 40, 41, 42, 43) - The mixed GRM (44, 45, 46) - The mixed NRM (47) - The two-dimensional mixed PCM (48)	(+) No additional items are necessary. (+) Models are applicable to categorical data. (+) Models can detect diverse RSs in the data in an exploratory way, with post hoc interpretation of RSs from class-specific item parameter estimates. (+) Individual latent trait values that are adjusted for RSs are provided by applying the model. (–) RSs are assumed as categorical variables.

Statistical approach	Description and representative models	Advantages and disadvantages
Mixture IRT approach with random-effect factor	Extended mixed IRT model by including a random-effect to quantify the ERS. These models help detect latent classes with different response patterns and quantify an individual tendency for the ERS. The additional random-effect factor represents interindividual differences in category widths. Examples: - The mixture ERS-GPCM [(49)] - The mixture ERS-GPCM with an additional item-specific constrained discrimination parameter. This model also makes it possible to identify items that strongly evoke ERS [(49).]	(+) No additional items are necessary. (+) Calculating the ERS intensity occurs from distances between adjacent thresholds. Small distances indicate an individual tendency to ERS use. (+) Classifying respondents to latent classes with different RSs and quantifying individual tendency for the ERS are simultaneously possible. (–) Models account only for scale-specific ERS. (–) Respondents with high trait levels cannot be separated from respondents with high ERS levels.

Notes. RS = Response style; ANCOVA = Analysis of covariance; CFA = Confirmatory factor analysis; IRT = Items response theory; ERS = Extreme response style; MRS = Middle response style; ARS = Acquiescence response style; DRS = Disacquiescence response style. NARS = Net acquiescence response style; GRM = Graded response model. PCM = Partial credit model. FA = Factor analysis; NRM = Nominal response model. SDR = socially desirable responses; IR = Item response. LCFA = Latent-class confirmatory factor analysis.

(+) stands for an advantage. (–) stands for a disadvantage.

[1] Bachman & O'Malley (1984); [2] Hui & Triandis (1985); [3] Baumgartner & Steenkamp (2001); [4] Weijters, Schillewaert, & Geuens (2008); [5] Greenleaf (1992a); [6] Greenleaf (1992b); [7] Reynolds & Smith (2010); [8] Thomas, Abts, & Vander Weyden (2014); [9] Weijters Cabooter, & Schillewaert (2010); [10] Weijters, Geuens, & Schillewaert (2010a); [11] Wetzel & Carstensen (2017); [12] He & van de Vijver (2013); [13] Billiet & McClendon (2000); [14] Johnson (2003); [15] Rossi, Gilula, & Allenby (2001); [16] Wang, Wilson, & Shih (2006); [17] Wang & Wu (2011); [18] Jin & Wang (2014); [19] Ferrando (2014); [20] Bolt & Johnson (2009); [21] Bolt & Newton (2011); [22] Falk & Cai (2015); [23] Bolt, Lu, & Kim (2014); [24] Javaras & Ripley (2007); [25] Mõttus, Allik, Realo, Rossier, Zecca, Ah-Kion, ... & Johnson (2012); [26] He, Van de Vijver, Espinosa, & Mui (2014); [27] Thissen-Roe & Thissen (2013); [28] Böckenholt (2012); [29] Plieninger & Meiser (2014); [30] Jeon & De Boeck (2015); [31] De Boeck & Partchev (2012); [32] Kieruj & Moors (2013); [33] Moors (2003); [34] Moors (2012); [35] Moors, Kieruj, & Vermunt (2014); [36] Morren, Gelissen, & Vermunt (2012); [37] Van Rosmalen, Van Herk, & Groenen (2010); [38] Meiser & Machunsky (2008); [39] Eid & Rauber (2000); [40] Jasper, Nater, Hiller, Ehlert, Fischer, & Witthöft (2013); [41] Rost, Carstensen, & von Davier (1999); [42] Wetzel, Böhnke, Carstensen, Ziegler, & Ostendorf (2013); [43] Wu & Huang (2010); [44] Egberink, Meijer, & Veldkamp (2010); [45] Gnambs & Hanfstingl (2014); [46] Sawatzky, Ratner, Kopec, & Zumbo (2012); [47] Maij-de Meij, Kelderman, & van der Flier (2008); [48] Böckenholt & Meiser (2017); [49] Huang (2016).

latent classes can be interpreted from estimated class-specific item parameters. The mixed IRT models are a promising approach that has proven suitable for identifying different types of ICU, such as RSs (Austin et al., 2006; Eid & Rauber, 2000; Gnambs & Hanfstingl, 2014; Meiser & Machunsky, 2008; Wetzel, Böhnke, Carstensen, Ziegler, & Ostendorf, 2013; Wu & Huang, 2010), socially desirable responses (Eid & Zickar, 2007; Maij-de Meij et al., 2008; Mneimneh, Heeringa, Tourangeau, & Elliott, 2014; Zickar, Gibby, & Robie, 2004; Ziegler & Kemper, 2013), and specific shortcut strategies (e.g., Sawatzky, Ratner, Kopec, & Zumbo, 2012). The number of latent classes and response behavior within latent classes can vary depending on the trait investigated, the rating scale used, and the sample recruited. Furthermore, an application of the model provides respondents' trait values that are adjusted for ICU (Rost, 1997). Therefore, researchers can use the estimated person parameters for subsequent analyses.

However, less evidence is present on how mixed polytomous IRT models work when applied to data originating from national surveys in which short scales with many response categories are standardly used for data collection. For instance, it is unclear how large the sample should be to obtain accurate model parameter estimates. It is also questionable what information criteria are optimal for identifying the best-fitting class solution. Due to their relevance, these issues are also examined in this dissertation.

1.7 The Present Dissertation

The present dissertation is based on three separate studies. Two were empirical studies that investigated the adequacy of the 11-point rating scale as a valid assessment of job satisfaction (Chapter 2) and whether a shorter rating scale would be more optimal for this purpose (Chapter 4). In these studies, the occurrence of ICU served as the main evaluation criterion. Therefore, the mixed polytomous IRT models were applied in both studies to detect patterns of ICU in an exploratory way. Furthermore, these two studies differed in their methodological approach. In Chapter 2, data from a representative panel survey (HILDA survey) were analyzed. In Chapter 4, an experimental design was applied by testing an effect of rating scale length on ICU. For this purpose, the data from an online sample were collected. In addition, both studies examine the personal profiles of RS users. In addition, the Monte Carlo simulation study (Chapter 3), which was conducted at the preparation stage of the experimental study, addressed the performance of mixed polytomous IRT models when applied on data measured with a long rating scale. In the following section, the specific aims and methods of each study are presented in more detail.

1.7.1 Study of Adequacy of a Long Rating Scale (Chapter 2)

The goal of the study presented in Chapter 2 was to gather empirical evidence on the limited adequacy of an 11-point rating scale due to ICU. This response format is considered standard in national panel surveys for assessing cognitive well-being. This study focuses on items of job satisfaction. If a tolerable extent of ICU is present in the data, this evidence would suggest that many response categories validly assess fine-grained differences between respondents in their levels of job satisfaction. If, on the contrary, a high extent of ICU exists in the data, this would indicate that this long rating scale overstrains respondents with superfluous categories and, therefore, provokes biased responses. A further goal of this study was to describe the profile of respondents who are inclined to RSs.

To determine the adequacy of an 11-point rating scale, a methodological approach was required to detect patterns of category use. Chapter 2 considers data on aspects of job satisfaction (5 items) from a representative sample of employees and employers provided by the HILDA survey (*n* = 7,036). This data is analyzed with the mixed polytomous IRT models (von Davier & Carstensen, 2007). These models provided the ability to identify latent classes differing in patterns of category use from estimated item parameters, without requiring us to define any types of ICU a priori. By applying these models, we were able to detect what RSs are prevalent in the sample and what number of categories are avoided due to their redundancy. The study presented in Chapter 2 is the first to examine the adequacy of a long rating scale (11 points) in the context of national panel studies.

1.7.2 Simulation Study (Chapter 3)

A few simulation studies have examined the performance of mixed polytomous IRT models. Moreover, most of them have included perfect data conditions (e.g., a large number of items in a scale, a rating scale with a few response categories), which hardly conform to data situations of national panel surveys that assess cognitive well-being. The goal of the Monte Carlo simulation study presented in Chapter 3 was to examine the performance of mixed polytomous IRT models when applied to the data assessed with a few items and a long rating scale, as is the case in national panel surveys. This simulation study focused on two models, the restricted mixed generalized partial credit model (rmGPCM; mGPCM; von Davier & Yamamoto, 2004; GPCM; Muraki, 1997) and the mixed partial credit model (mPCM; Rost, 1997). Both models were established for detecting patterns of category use in an exploratory way. The simulation design was guided to answer two research questions: (1) What sample size is required for the appropriate performance of these mixed polytomous IRT models? (2) What information criteria are effective for model selection? A particular feature of this simulation study was that the data was generated based on model parameters empirically derived from an application of these models in the study reported in Chapter 2. When applied to problematic data, the population parameters of the two models represent the realistic features of category use and, therefore, ensure the ecological validity of the simulation study. Notably, this simulation study was conducted at the preparation stage of the experimental study (Chapter 4) to determine how many respondents should be recruited within each experimental condition for the proper performance of mixed polytomous IRT models.

1.7.3 Experimental Study (Chapter 4)

Almost all previous experimental studies examining the effects of rating scale length on ICU focused on a few traditional RSs. The goal of the experimental study presented in Chapter 4 was to identify an optimal rating scale for assessing job satisfaction by exploring patterns of category use in three experimental conditions, wherein the number of response categories is varied (long rating scale with 11 categories and two short rating scales with six and four options, respectively). Two hypotheses were proposed: (1) the short rating scales evoke less ICU (e.g., fewer avoided response categories), indicating that respondents can effectively cope with shorter rating scales; and (2) optimizing rating scale length has hardly any impact on the ERS, suggesting that this RS occurs due to specific dispositions or respondents. In Chapter 4, data from a sample of American employees and employers who are registered on Amazon's Mechanical Turk (MTurk) platform (*N* = 6,999) was collected by randomly assigning them to one of three experimental conditions. Within each of these experimental conditions, category use was analyzed using the multidimensional version of the rmGPCM. An advantage of this model was that it allowed us to simultaneously identify an invariant structure of latent classes with homogeneous patterns of category use across all dimensions of job satisfaction. This latent mixture makes sense because of previous evidence suggesting the consistent category use across different subscales. Moreover, this study allowed us to systematically examine whether potential predictors such as socio-demographic variables, personality traits, cognitive ability, and job-related variables account for consistent use of a particular RS across experimental conditions. Hence, the study presented in this dissertation provides insight into the potential of optimizing the rating scale and profiles of consistent RS users.

1.7.4 General Discussion (Chapter 5)

In Chapter 5, the findings from all three studies are integrated and discussed, with special emphasis on generalizability and practical implications. This dissertation concludes with suggestions for future studies.

1.8 References

Aichholzer, J. (2014). Random intercept EFA of personality scales. *Journal of Research in Personality, 53*, 1-4. http://dx.doi.org/10.1016/j.jrp.2014.07.001.

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, NY: Routledge.

Arce-Ferrer, A. J. (2006). An Investigation Into the Factors Influencing Extreme-Response Style. *Educational and Psychological Measurement, 66*(3), 374-392.

Ashton, M. C., & Lee, K. (2007). Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. *Personality and Social Psychology Review, 11*(2), 150–166.

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*, 1235-1245. doi: 10.1016/j.paid.2005.10.018

Bachman, J. G., & O'Malley, P. M. (1984). Yea-saying, nay-saying, and going to extremes: Black-white differences in response styles. *Public Opinion Quarterly, 48*(2), 491-509. doi: 10.1086/268845

Baumgartner, H., & Steenkamp, J.-B. E. M. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research, 38*, 143-156. doi: 10.1509/jmkr.38.2.143.18840

Berg, I. A., & Collier, J. S. (1953). Personality and group differences in extreme response sets. *Educational and Psychological Measurement, 13*(2), 164-169. doi: 10.1177/001316445301300202

Billiet, J. B., & Davidov, E. (2000). Testing the stability of an acquiescence style factor behind two interrelated substantive variables in a panel design. *Sociological Methods & Research, 36*(4), 542–562. doi: 10.1177/0049124107313901

Billiet, J. B., & McClendon, M. J. (2000). Modeling acquiescence in measurement models for two balanced sets of items. *Structural Equation Modeling*, *7*(4), 608-628. doi: 10.1207/S15328007SEM0704_5

Böckenholt, U. (2012). Modeling multiple response processes in judgment and choice. *Psychological Methods, 17*(4), 665-678. doi: 10.1037/a0028111

Böckenholt, U., & Meiser, T. (2017). Response style analysis with threshold and multi-process IRT models: A review and tutorial. *British Journal of Mathematical and Statistical Psychology*, *70*(1), 159-181. https://doi.org/10.1111/bmsp.12086

Bolt, D. M., & Johnson, T. R. (2009). Addressing score bias and differential item functioning due to individual differences in response style. *Applied Psychological Measurement*, *33*(5), 335-352. https://doi.org/10.1177/0146621608329891

Bolt, D. M., & Newton, J. R. (2011). Multiscale measurement of extreme response style. *Educational and Psychological Measurement*, *71*(5), 814-833. https://doi.org/10.1177/0013164410388411

Bolt, D. M., Lu, Y., & Kim, J. S. (2014). Measurement and control of response styles using anchoring vignettes: A model-based approach. *Psychological Methods*, *19*(4), 528-541. doi:10.1037/met0000016

Börsch-Supan, A. (2013). *SHARE—Survey of Health, Ageing and Retirement in Europe.* Available online at http://www.share-project.org [Accessed 30 November 2018].

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. doi: 10.1348/096317909x478557

Brengelman, J. (1960). Extreme response set, drive level and abnormality in questionnaire rigidity. *Journal of Mental Science, 106*, 171–186. doi: 10.1192/bjp.106.442.171.

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. doi: 10.1007/s11135-016-0411-5

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. doi: 10.1016/j.jbusres.2015.10.138

Cabrita, J., & Perista, H. (2007). *Measuring job satisfaction in surveys–comparative analytical report.* European Foundation for the Improvement of Living and Working Condition. Available online at https://www.eurofound.europa.eu/publications/2006/working-conditions/measuring-job-satisfaction-in-surveys-comparative-analytical-report [Accessed 30 November 2018].

Cacioppo, J. T, & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology, 42*(1), 116-131. doi: 10.1037/0022-3514.42.1.116

Cacioppo, J. T., Petty, R. E., & Feng Kao, C. (1984). The efficient assessment of need for cognition. *Journal of personality assessment*, *48*(3), 306-307.

Calsyn, R. J., Roades, L. A., & Calsyn, D. S. (1992). Acquiescence in needs assessment studies of the elderly. *The Gerontologist*, *32*(2), 246-252.

Cammann, C., Fichman, M., Jenkins, G. D., & Klesh, J. (1983). Michigan Organizational Assessment Questionnaire. In S. E. Seashore, E. E. Lawler, P. H. Mirvis, & C. Cammann (Eds.), *Assessing organizational change: A guide to methods, measures, and practices* (pp. 71–138). New York: Wiley-Interscience.

Carr, L. G. (1971). Srole items and acquiescence. *American Sociological Review, 36*(2), 287-293. doi: 10.2307/2094045

Carter, N. T., Dalal, D. K., Lake, C. J., Lin, B. C., & Zickar, M. J. (2011). Using mixed-model item response theory to analyze organizational survey responses: An illustration using the Job Descriptive Index. *Organizational Research Methods, 14*, 116-146. doi:10.1177/1094428110363309

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. doi: 10.1177/014662169401800302

Cheung, G. W., & Rensvold, R. B. (2000). Assessing extreme and acquiescence response sets in cross-cultural research using structural equations modeling. *Journal of Cross-Cultural Psychology, 31*(2), 187-212.

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. doi: 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. doi: 10.2307/3150495

Cronbach, L. J. (1950). Further evidence on response sets and test design. *Educational and Psychological Measurement*, *10*(1), 3-31.

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. doi: 10.1177/0146621612470210

Cummins, R. A., & Gullone, E. (2000). Why we should not use 5-point Likert scales: The case for subjective quality of life measurements. In *Proceedings, Second International Conference on Quality of Life in Cities*, (pp. 74-93). Singapore: National University of Singapore.

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, *66*, 4-19. doi: 10.1016/j.jesp.2015.07.006

Danner, D., Aichholzer, J., & Rammstedt, B. (2015). Acquiescence in personality questionnaires: Relevance, domain specificity, and stability. *Journal of Research in Personality*, *57*, 119-130. doi: 10.1016/j.jrp.2015.05.004

Das, J. P., & Dutta, T. (1969). Some correlates of extreme response set. *Acta Psychologica, 29*(1), 85-92. doi: 10.1016/0001-6918(69)90005-5

De Boeck, P., & Partchev, I. (2012). IRTrees: Tree-based item response models of the GLMM family. *Journal of Statistical Software*, *48*(1), 1-28. http://www.jstatsoft.org/

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*, 104–115. doi: 10.1509/jmkr.45.1.104

DeVellis, R. F. (2016). *Scale development: Theory and applications* (Vol. 26). Sage publications.

Diener, E. D., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The satisfaction with life scale. *Journal of Personality Assessment*, *49*(1), 71-75.

Diener, E., Lucas, R. E., & Oishi, S. (2002). Subjective well-being: The science of happiness and life satisfaction. In C. R. Snyder & S. J. Lopez (Eds.), *The handbook of positive psychology* (pp. 63–73). Oxford, England: Oxford University Press.

Diener, E., Lucas, R., Schimmack, U., & Helliwell, J. (2009). *Well-being for public policy*. USA: Oxford University Press.

Diener, E., & Ryan, K. (2009). Subjective well-being: A general overview. *South African Journal of Psychology*, *39*(4), 391-406.

Diener, E., & Suh, E. (1997). Measuring quality of life: Economic, social, and subjective indicators. *Social indicators research*, *40*(1-2), 189-216.

Diener, E., & Suh, E. M. (2000). Measuring subjective well-being to compare the quality of life of cultures. In E. Diener and E. M. Suh (eds.), *Culture and Subjective Well-Being* (pp. 3–12). Cambridge, MA: The MIT Press.

Diener, E., Inglehart, R., & Tay, L. (2013). Theory and validity of life satisfaction scales. *Social Indicators Research*, *112*(3), 497-527.

Diener, E., Scollon, C. N., & Lucas, R. E. (2009). The evolving concept of subjective well-being: The multifaceted nature of happiness. In D. Ed (Ed.), *Assessing well-being: The collected works of Ed Diener*. Social Indicators Research Series, Volume 39 (pp. 67-100). Dordrecht: the Netherlands, Springer.

Dittmann-Kohli, F., Kohli, M., Künemund, H., Motel, A., Steinleitner, C., & Westerhof, G. J. (1997). *Lebenszusammenhänge, Selbst- und Lebenskonzeptionen. Erhebungsdesign und Instrumente des Alters-Survey* [*Life Associations, Self and Life Conceptions: Methodological Design of the German Ageing Survey*]. Forschungsgruppe Altern und Lebenslauf (FALL), Freie Universität Berlin, Forschungsbericht 47. Available online at http://www.fall-berlin.de/ [Accessed 30 November 2018].

Dolnicar, S., & Grün, B. (2009). Does one size fit all? The suitability of answer formats for different constructs measured. *Australasian Marketing Journal, 17*(1), 58-64.

Dubois, B., & Burns, J. A. (1975). An analysis of the meaning of the question mark response category in attitude scales. *Educational and Psychological Measurement*, *35*(4), 869-884.

Duffy, D., Leissou, E., McGonagle, K., & Schlegel, J. (2013). *PSID Main Interview User Manual: Release 2013.* Institute for Social Research, University of Michigan.

Egberink, I. J., Meijer, R. R., & Veldkamp, B. P. (2010). Conscientiousness in the workplace: Applying mixture IRT to investigate scalability and predictive validity. *Journal of Research in Personality, 44*(2), 232-244.

Eid, M. (2008). Measuring the immeasurable: Psychometric modeling of subjective well-being data. In M. Eid & R. J. Larsen (Eds.), The science of subjective well-being. New York: Guilford.

Eid, M., & Rauber, M. (2000). Detecting measurement invariance in organizational surveys. *European Journal of Psychological Assessment, 16*(1), 20-30. doi: 10.1027//1015-5759.16.1.20

Eid, M., & Zickar, M. J. (2007). Detecting response styles and faking in personality and organizational assessments by mixed Rasch models. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models* (pp. 255-270). New York, NY: Springer

Elliott, L. L. (1961). Effects of item construction and respondent aptitude on response acquiescence. *Educational and Psychological Measurement, 21*(2), 405–415.

ESS Data Team (2017). *ESS8 2016 Data Protocol.* Norwegian Centre for Research Data. Available online at https://www.europeansocialsurvey.org/data/ Accessed 30 November 2018].

EuroPanel Users Network (2004). *ECHP User Guide.* Institute for Social and Economic Research, University of Essex, Available online at http://epunet.essex.ac.uk/ECHP_USER_GUIDE_28-11-2005.pdf [Accessed 30 November 2018].

European Foundation for the Improvement of Living and Working Conditions (2010). *Second European quality of life survey.* Subjective well-being in Europe. Dublin: European Foundation for the Improvement of Living and Working Conditions.

Falk, C. F., & Cai, L. (2015). A flexible full-information approach to the modeling of response styles. *Psychological Methods, 21*(3), 328{347. doi:10.1037/met0000059

Faragher, E. B., Cass, M., & Cooper, C. L. (2005). The relationship between job satisfaction and health: A meta-analysis. *Occupational Environmental Medicine, 62*, 105–112. doi: 10.1136/oem.2002.006734

Ferrando, P. J. (2014). A factor-analytic model for assessing individual differences in response scale usage. *Multivariate Behavioral Research*, *49*(4), 390-405. https://doi.org/10.1080/00273171.2014.911074

Forehand, G. A. (1962). Relationships among response sets and cognitive behaviors. *Educational and Psychological Measurement, 22*(2), 287-302. doi: 10.1177/001316446202200204

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.

Frick, J. R., Jenkings, S. P., Lillard, D. R., Lipps, O., & Wooden, M. (2007). The Cross-National Equivalent File (CNEF) and its member country household panel studies. *Schmollers Jahrbuch: Zeitschrift für Wirtschafts-und Sozialwissenschaften*, *127*(4), 627-654.

Gerber-Braun, B. (2010). *The Double Cross: Individual differences between respondents with different response sets and styles on questionnaires.* Dissertation, Ludwig–Maximilians–Universität, München.

Gibbons, J. L., Zellner, J. A., & Rudek, D. J. (1999). Effects of language and meaningfulness on the use of extreme response style by Spanish-English bilinguals. *Cross-Cultural Research*, *33*(4), 369-381.

Glatzer, W., Camfield, L., Møller, V., & Rojas, M. (Eds.). (2015). *Global handbook of quality of life: Exploration of well-being of nations and continents.* Springer.

Gnambs, T., & Hanfstingl, B. (2014). A differential item functioning analysis of the German academic self-regulation questionnaire for adolescents. European *Journal of Psychological Assessment, 30*(4), 251-260. doi: 10.1027/1015-5759/a000185

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. doi: 10.2307/3172568

Greenleaf, E. A. (1992b). Measuring extreme response style. *Public Opinion Quarterly, 56*(3), 328-351. doi: 10.1086/269326

Gudjonsson, G. H. (1986). The relationship between interrogative suggestibility and acquiescence: Empirical findings and theoretical implications. *Personality and Individual Differences*, *7*(2), 195-199.

Gudjonsson, G. H. (1990). The relationship of intellectual skills to suggestibility, compliance and acquiescence. *Personality and Individual Differences, 11*(3), 227-231. doi: 10.1016/0191-8869(90)90236-K

Gudjonsson, G. H., & Clare, I. C. (1995). The relationship between confabulation and intellectual ability, memory, interrogative suggestibility and acquiescence. *Personality and Individual Differences*, *19*(3), 333-338.

Hamby, T., & Levine, D. S. (2016). Response-scale formats and psychological distances between categories. *Applied Psychological Measurement*, *40*(1), 73-75. doi: 10.1177/0146621615597961

Hamilton, D. (1968). Personality attributes associated with extreme response style. *Psychological Bulletin, 69*(3), 192–203. doi: 10.1037/h0025606.

Harumi, C. A. (2011). *Cross-cultural differences in response styles*. Doctoral dissertation, Washington State University.

Harzing, A. W., Baldueza, J., Barner-Rasmussen, W., Barzantny, C., Canabal, A., Davila, A., ... & Liang, Y. K. (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. doi: 10.1016/j.ibusrev.2009.03.001

He, J., & van de Vijver, F. J. (2013). A general response style factor: Evidence from a multi-ethnic study in the Netherlands. *Personality and Individual Differences*, *55*(7), 794-800. https://doi.org/10.1016/j.paid.2013.06.017

He, J., Van de Vijver, F. J., Espinosa, A. D., & Mui, P. H. (2014). Toward a unification of acquiescent, extreme, and midpoint response styles: A multilevel study. *International Journal of Cross Cultural Management*, *14*(3), 306-322. https://doi.org/10.1177/1470595814541424

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. doi: 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*, 1706. doi: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. doi: 10.1177/0022022189203004

Inglehart, R., C. Haerpfer, A. Moreno, C. Welzel, K. Kizilova, J. Diez-Medrano, M. Lagos, P. Norris, E. Ponarin & B. Puranen et al. (2014). *World Values Survey: Round Six - Country-Pooled Datafile*

Version. Madrid: JD Systems Institute. Available online at www.worldvaluessurvey.org/WVSDocumentationWV6.jsp. [Accessed 30 November 2018].

Jasper, F., Nater, U. M., Hiller, W., Ehlert, U., Fischer, S., & Witthöft, M. (2013). Rasch scalability of the somatosensory amplification scale: a mixture distribution approach. *Journal of Psychosomatic Research*, *74*(6), 469-478. https://doi.org/10.1016/j.jpsychores.2013.02.006

Javaras, K. N., & Ripley, B. D. (2007). An "unfolding" latent variable model for Likert attitude data: Drawing inferences adjusted for response style. *Journal of the American Statistical Association*, *102*(478), 454-463. doi: 10.1198/016214506000000960

Jeon, M., & De Boeck, P. (2016). A generalized item response tree model for psychological assessments. *Behavior Research Methods*, *48*(3), 1070-1085. https://doi.org/10.3758/s13428-015-0631-y

Jin, K. Y., & Wang, W. C. (2014). Generalized IRT models for extreme response style. *Educational and Psychological Measurement, 74*, 116-138. doi: 10.1177/0013164413498876

Johnson, T. (2003). On the use of heterogeneous thresholds ordinal regression models to account for individual differences in response style. *Psychometrika, 68*(4), 563-583. https://doi.org/10.1007/BF02295612

Johnson, T., Kulesa, P., Cho, Y. I., & Shavitt, S. (2005). The relation between culture and response styles: Evidence from 19 countries. *Journal of Cross-Cultural Psychology, 36*(2), 264–277. http://dx.doi.org/10.1177/0022022104272905.

Judge, T. A., Thorensen, 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. doi: 10.1037/0033-2909.127.3.376

Keller, F., & Koller, I. (2015). Mixed Rasch Models for Analyzing the Stability of Response Styles Across Time: An Illustration with the Beck Depression Inventory (BDI-II). In M. Stemmler, A. von Eye, & W. Wiedermann (Eds.), *Dependent Data in Social Sciences Research* (pp. 309-324). Springer, Cham.

Kerrick, J. S. (1954). *The effects of intelligence and manifest anxiety on attitude change through communications.* Doctoral dissertation, University of Illinois at Urbana-Champaign. http://hdl.handle.net/2142/58568

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. doi: 10.1186/1477-7525-10-80

Khorramdel, L., & von Davier, M. (2014). Measuring Response Styles Across the Big Five: A Multiscale Extension of an Approach Using Multinomial Processing Trees. *Multivariate Behavioral Research, 49*(2), 161-177.

Kieruj, N. D. (2012). *Question format and response style behavior in attitude research*. Uitgeverij BOXPress.

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. doi: 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. doi: 10.1007/s11135-011-9511-4

Knies, G. (2017). *Understanding society: The UK household longitudinal study, Waves 1–7 (User Guide).* Colchester: Institute for Social and Economic Research, University of Essex.

Knowles, E. S. & Nathan K. T. (1997). Acquiescent Responding in Self-Reports: Cognitive Style or Social Concern. *Journal of Research in Personality, 31*(2), 293-301. doi: 10.1006/jrpe.1997.2180

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology, 5*(3), 213-236. doi: 10.1002/acp.2350050305

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology, 50*, 537-567. doi: 10.1146/annurev.psych.50.1.537

Krosnick, J. A., & Alwin, D. F. (1987). An evaluation of a cognitive theory of response-order effects in survey measurement. *Public Opinion Quarterly*, *51*(2), 201-219. doi: 10.1086/269029

Krosnick, J. A., & Presser, S. (2010). Question and questionnaire design. In J. D. Wright & P. V. Marsden (Eds.), *Handbook of survey research* (pp. 263–314). Bingley: Emerald Group Publishing Ltd.

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. doi: 10.1016/j.jrp.2013.01.014

Lau, M. Y. (2007). *Extreme Response Style: An Empirical Investigation of the Effects of Scale Response Format Fatigue.* PhD dissertation, University of Notre Dame, Notre Dame, IN.

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. doi: 10.1177/0734282914522200

Lewis, N. A., & Taylor, J. A. (1955). Anxiety and extreme response preferences. *Educational and Psychological Measurement*, *15*(2), 111-116.

Light, C. S., Zax, M., & Gardiner, D. H. (1965). Relationship of age, sex, and intelligence level to extreme response style. *Journal of Personality and Social Psychology, 2*(6), 907-909.

Liu, M., & Keusch, F. (2017). Effects of scale direction on response style of ordinal rating scales. *Journal of Official Statistics*, *33*(1), 137-154. doi: 10.1515/jos-2017-0008

Lozano, L. M., García-Cueto, E., and Muñiz, J. (2008). Effect of the number of response categories on the reliability and validity of rating scales. *Methodology, 4*(2), 73-79. doi: 10.1027/1614-2241.4.2.73.

Lu, Y., & Bolt, D. M. (2015). Examining the attitude-achievement paradox in PISA using a multilevel multidimensional IRT model for extreme response style. *Large-Scale Assessments in Education*, *3*(1), 1-18. https://doi.org/10.1186/s40536-015-0012-0

Lucas, R. E., & Donnellan, M. B. (2012). Estimating the reli-ability of single-item life satisfaction measures: Results from fournational panel studies. *Social Indicators Research, 105*(3), 323–331. doi: 10.1007/s11205-011-9783-z

Luhmann, M., Hofmann, W., Eid, M., & Lucas, R. E. (2012). Subjective well-being and adaptation to life events: A meta-analysis. *Journal of Personality and Social Psychology*, *102*(3), 592-615. doi: 10.1037/a0025948

Lyubomirsky, S., King, L., & Diener, E. (2005). The benefits of frequent positive affect: Does happiness lead to success? *Psychological Bulletin, 131*, 803-855.

Maij-de Meij, A. M., Kelderman, H., & van der Flier, H. (2008). Fitting a mixture item response theory model to personality questionnaire data: Characterizing latent classes and investigating possibilities for improving prediction. *Applied Psychological Measurement, 32*, 611-631. doi:10.1177/0146621607312613

Marchand, M. (2017). *Personality - LISS (Longitudinal Internet Studies for the Social Sciences) Core Study: Questionnaire administered to the LISS panel (Wave 9).* Institute for Data Collection and Research

(CentERdata), Tilburg University, The Netherlands. Available online at https://www.dataarchive.lissdata.nl/study_units/view/668 [Accessed 30 November 2018].

Marín, G., Gamba, R. J., & Marín, B. V. (1992). Extreme response style and acquiescence among Hispanics: The role of acculturation and education. *Journal of Cross-Cultural Psychology, 23*(4), 498-509. doi: 10.1177/0022022192234006

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*(2), 295-308.

McClendon, M. J. (1991). Acquiescence and recency response-order effects in interview surveys. *Sociological Methods & Research*, *20*(1), 60-103. doi: 10.3758/BRM.41.2.295

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437-455. doi: 10.1037/a0028085

Meisenberg, G., & Williams, A. (2008). Are acquiescent and extreme response styles related to low intelligence and education? *Personality and Individual Differences, 44*(7), 1539–1550. doi: 10.1016/j.paid.2008.01.010.

Meiser, T., & Machunsky, M. (2008). The personal structure of personal need for structure. *European Journal of Psychological Assessment*, *24*(1), 27-34. doi: 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*, 81–97. doi: 10.1037/h0043158

Mirowsky, J., & Ross, C. E. (1991). Eliminating defense and agreement bias from measures of the sense of control: A 2 x 2 index. *Social Psychology Quarterly*, *54*, 127-145. doi: 10.2307/2786931

Mneimneh, Z. N., Heeringa, S. G., Tourangeau, R., & Elliott, M. R. (2014). Bridging psychometrics and survey methodology: Can mixed Rasch models identify socially desirable reporting behavior?. *Journal of Survey Statistics and Methodology*, *2*(3), 257-282. doi: 1093/jssam/smu008

Moors, G. (2003). Diagnosing response style behavior by means of a latent-class factor approach. Socio-demographic correlates of gender role attitudes and perceptions of ethnic discrimination reexamined. *Quality & Quantity, 37*(3), 277-302. doi:10.1023/A:1024472110002

Moors, G. (2008). Exploring the effect of a middle response category on response style in attitude measurement. *Quality & Quantity*, *42(6)*, 779-794. doi:10.1007/s11135-006-9067-x

Moors, G. (2012). The effect of response style bias on the measurement of transformational, transactional, and laissez-faire leadership. *European Journal of Work and Organizational Psychology*, *21*(2), 271-298.

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. doi: 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.

Mõttus, R., Allik, J., Realo, A., Rossier, J., Zecca, G., Ah-Kion, J., ... & Bhowon, U. (2012). The effect of response style on self-reported conscientiousness across 20 countries. *Personality and Social Psychology Bulletin*, *38*(11), 1423-1436.

Muraki, E. (1997). A generalized partial credit model. In W. van der Linden & R. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 153-164). New York: Springer.

Naemi, B. D., Beal, D. J., & Payne, S. C. (2009). Personality predictors of extreme response style. *Journal of Personality, 77*(1), 261-286. doi: 10.1111/j.1467-6494.2008.00545.x

NatCen Social Research (2012). *English Longitudinal Study of Ageing (ELSA): Wave One to Wave Five. User Guide to the datasets.* NatCet Social Research that Works for Society. Available online at https://www.elsa-project.ac.uk/publicationDetails/id/6791 [Accessed 30 November 2018].

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

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology*, *46*(3), 598-609.

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of Personality and Social Psychological Attitudes* (17-59). San Diego, CA US: Academic Press.

Pedersen, D. M. (1967). Acquiescence and social desirability response sets and some personality correlates. *Educational and Psychological Measurement*, *27*(3), 691-697.

Plieninger, H., & Meiser, T. (2014). Validity of multiprocess IRT models for separating content and response styles. *Educational and Psychological Measurement, 74*(5), 875-899. doi:10.1177/0013164413514998

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. doi: 10.1037/0021-9010.88.5.879

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. doi:10.1016/S0001-6918(99)00050-5

Rafferty, A. E., & Griffin, M. A. (2009). Job satisfaction in organizational research. In D. A. Buchanan & A. Bryman (Eds.), *The Sage handbook of organizational research methods* (pp. 196–212). London, UK: Sage.

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. doi: 10.1177/0049124113509605

Reynolds, N., & Smith, A. (2010). Assessing the impact of response styles on cross-cultural service quality evaluation: a simplified approach to eliminating the problem. *Journal of Service Research, 13*(2), 230-243.

Richter, D., Rohrer, J., Metzing, M., Nestler, W., Weinhardt, M., & Schupp, J. (2017). *SOEP Scales Manual (updated for SOEP-Core v32.1)*. SOEP Survey Papers 423: Series C. Berlin: DIW/SOEP.

Ross, C. E., & Mirowsky, J. (1984). Socially-desirable response and acquiescence in a cross-cultural survey of mental health. *Journal of Health Social Behavior, 25*(2), 189-197. doi: 10.2307/2136668

Ross, C. K., Steward, C. A., & Sinacore, J. M. (1995). A comparative study of seven measures of patient satisfaction. *Medical Care, 33*(4), 392–406. doi: 10.1097/00005650-199504000-00006

Rossi, P. E., Gilula, Z., & Allenby, G. M. (2001). Overcoming Scale Usage Heterogeneity. *Journal of the American Statistical Association, 96*(453), 20-31. doi: 10.1198/016214501750332668

Rost, J. (1997). Logistic mixture models. In W. van der Linden & R. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 449-463). New York: Springer.

Rost, J., Carstensen, C. H., & von Davier, M. (1999). Sind die Big Five Rasch-skalierbar? Eine Reanalyse der NEO-FFI-Normierungsdaten [Are the Big Five Rasch scaleable? A reanalysis of the NEO-FFI norm data]. *Diagnostica, 45*(3), 119-127. http://dx.doi.org/10.1026//0012-1924.45.3.119

Ryff, C., Almeida, D. M., Ayanian, J., Carr, D. S., Cleary, P. D., Coe, C., & Williams, D. (2017). *Midlife in the United States (MIDUS 2), 2004-2006*. Ann Arbor, MI: Inter-university Consortium for Political and Social Research. doi:10.3886/ICPSR04652.v7

Sawatzky, R., Ratner, P. A., Kopec, J. A., & Zumbo, B. D. (2012). Latent variable mixture models: a promising approach for the validation of patient reported outcomes. *Quality of Life Research, 21*(4), 637-650. doi: 10.1007/s11136-011-9976-6

Scherpenzeel, A. C., & Saris, W. E. (1997). The validity and reliability of survey questions: A meta-analysis of MTMM studies. *Sociological Methods and Research*, *25*(3), 341-383. doi: 10.1177/0049124197025003004

Schimmack, U., Krause, P., Wagner, G. G., & Schupp, J. (2010). Stability and change of well-being: An experimentally enhanced latent state-trait-error analysis. *Social Indicators Research, 95*, 19-31. doi: 10.1007/s11205-009-9443-8

Scholz, E., Jutz, R., Edlund, J., Öun, I., & Braun, M. (2014). *ISSP (International Social Survey Programme) 2012: Family and Changing Gender Roles IV. Questionnaire Development.* GESIS Leibnitz-Institut für Sozialwissenschaften. Available online at https://www.gesis.org/fileadmin/upload/forschung/publikationen/gesis_reihen/gesis_methodenberichte/2014/TechnicalReport_2014-11.pdf [Accessed 30 November 2018].

Schuman, H. & Presser, S. (1981). *Questions and Answers in Attitude Surveys*. New York: Academic Press.

Schwarz, N., Knäuper, B., Hippler, H. J., Noelle-Neumann, E., & Clark, L. (1991). Rating scales numeric values may change the meaning of scale labels. *Public Opinion Quarterly*, *55*(4), 570-582. doi: 10.1086/269282

Schwarz, N., Strack, F., Hippler, H. J., & Bishop, G. (1991). The impact of administration mode on response effects in survey measurement. *Applied Cognitive Psychology, 5*, 193–212. doi: 10.1002/acp.2350050304

Smith, P. C., Kendall, L., & Hulin, C. L. (1969). *The measurement of satisfaction in work and retirement: A strategy for the study of attitudes*. Chicago: Rand McNally.

Statistical Association (Eds.), *Proceedings of the survey research methods section* (pp. 1003–1012). Alexandria.

Streefkerk, M. (2017). *Work and Schooling – LISS (Longitudinal Internet Studies for the Social Sciences) Core Study: Questionnaire administered to the LISS panel (Wave 10).* Institute for Data Collection and Research (CentERdata), Tilburg University, The Netherlands. Available online at https://www.dataarchive.lissdata.nl/study_units/view/669 [Accessed 30 November 2018].

Summerfield, M., Bevitt, A., Freidin, S., Hahn, M., La, N., Macalalad, N., O'Shea, M., Watson, N., Wilkins, R. and Wooden, M. (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. doi: 10.1086/321952

Swedish Institute for Social Research (2010). *Level of Living Survey: English Codebook*. University Stockholm. Available online at https://www.sofi.su.se/english/2.17851/research/three-research-units/lnu-level-of-living [Accessed 30 November 2018].

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). Springer, Cham.

Tay, L., Diener, E., Drasgow, F., & Vermunt, J. K. (2011). Multilevel mixed-measurement IRT analysis: An explication and application to self-reported emotions across the world. *Organizational Research Methods*, *14*(1), 177-207. doi: 10.1177/1094428110372674

Taylor, M. F. (ed). with Brice, J., Buck, N., & Prentice-Lane, E. (2018). *British Household Panel Survey User Manual Volume A: Introduction, Technical Report and Appendices.* Colchester: University of Essex.

Thissen-Roe, A., & Thissen, D. (2013). A two-decision model for responses to likert-type items. *Journal of Educational and Behavioral Statistics, 38*(5), 522-547. https://doi.org/10.3102/1076998613481500

Thomas, T. D., Abts, K., & Vander Weyden, P. (2014). Measurement invariance, response styles, and rural–urban measurement comparability. *Journal of Cross-Cultural Psychology*, *45*(7), 1011-1027. https://doi.org/10.1177/0022022114532359

Tourangeau, R., & Smith, T. W. (1996). Asking sensitive questions: The impact of data collection mode, question format, and question context. *Public Opinion Quarterly*, *60*(2), 275-304. doi: 10.1086/297751

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response.* New York, NY US: Cambridge University Press.

Tutz, G., Schauberger, G., & Berger, M. (2018). Response styles in the partial credit model. *Applied Psychological Measurement, 42*(6), 407-427. doi: 10.1177/0146621617748322

Van Herk, H., Poortinga, Y. H., & Verhallen, T. M. M. (2004). Response styles in rating scales: Evidence of method bias in data from six EU countries. *Journal of Cross-Cultural Psychology, 35*(3), 346–360. doi: 10.1177/0022022104264126

Van Rosmalen, J., Van Herk, H., & Groenen, P. J. (2010). Identifying response styles: A latent-class bilinear multinomial logit model. *Journal of Marketing Research, 47*(1), 157-172. https://doi.org/10.1509/jmkr.47.1.157

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. doi: 10.1093/ijpor/eds021

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*, 108-124. doi:10.1016/s0148-2963(01)00296-x

von Davier, M., & Carstensen, C. H. (2007). *Multivariate and mixture distribution Rasch models*. New York: Springer.

von Davier, M., & Yamamoto, K. (2004). Partially observed mixtures of IRT models: An extension of the generalized partial-credit model. *Applied Psychological Measurement, 28*(6), 389-406. doi: 10.1177/0146621604268734

Voorpostel, M., Tillmann, R., Lebert, F., Kuhn, U., Lipps, O., Ryser, V. A., ... & Wernli, B. (2015). *Swiss Household Panel: User Guide (1999–2015). Wave*, *15*. Technical report, FORS.

Wagner-Menghin, M. M. (2006). The Mixed-Rasch Model: An Example for Analyzing the Meaning of Response Latencies in a Personality Questionnaire. *Journal of Applied Measurement*, *7*, 225-237.

Wagner, G. G., Frick, J. R., & Schupp, J. (2007). The German Socio-Economic Panel Study (SOEP). Scope, evolution and enhancements. *Schmollers Jahrbuch, 127*, 139-169.

Wakita, T., Ueshima, N., & Noguchi, H. (2012). Psychological distance between categories in the Likert scale: Comparing different numbers of options. *Educational and Psychological Measurement*, *72*(4), 533-546. doi: 10.1177/0013164411431162

Wang, W.-C., & Wu, S.-L. (2011). The random-effect generalized rating scale model. *Journal of Educational Measurement, 48*(4), 441-456. https://doi.org/10.1111/j.1745-3984.2011.00154.x

Wang, W.-C., Wilson, M., & Shih, C.-L. (2006). Modeling randomness in judging rating scales with a random-effects rating scale model. *Journal of Educational Measurement, 43*(4), 335-353. https://doi.org/10.1111/j.1745-3984.2006.00020.x

Warr, P. B., & Coffman, T. L. (1970). Personality, involvement and extremity of judgement. *British Journal of Social and Clinical Psychology*, *9*(2), 108-121. doi:10.1111/j.2044-8260.1970.tb00650.x

Watson, N., & Wooden, M. P. (2012). The HILDA survey: a case study in the design and development of a successful household panel survey. *Longitudinal and Life Course Studies*, *3*(3), 369-381. doi: 10.14301/llcs.v3i3.208

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*, 1516−1524. doi: 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. doi: 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. doi: 10.1177/0146621609338593

Weijters, B., Geuens, M., & Schillewaert, N. (2010b). The stability of individual response styles. *Psychological Methods, 15*(1), 96-110. doi: 10.1037/a0018721

Weijters, B., Schillewaert, N., & Geuens, M. (2008). Assessing response styles across modes of data collection. *Journal of the Academy of Marketing Science*, *36*(3), 409-422. doi: 10.1007/s11747-007-0077-6

Weiss, H. M. (2002). Deconstructing job satisfaction: Separating evaluations, beliefs and affective experiences. *Human Resource Management Review, 12*(2), 173–194. doi: 10.1016/S1053-4822(02)00045-1

Wetzel, E., & Carstensen, C. H. (2014). Reversed thresholds in partial credit models: A reason for collapsing categories?. *Assessment*, *21*(6), 765-774. doi: 10.1177/1073191114530775

Wetzel, E., & Carstensen, C. H. (2017). Multidimensional modeling of traits and response styles. *European Journal of Psychological Assessment, 33*(5), 352-364. doi: 10.1027/1015-5759/a000291

Wetzel, E., Böhnke, J. R., & Brown, A. (2016). Response biases. In F. T. L. Leong, D. Bartram, F. Cheung, K. F. Geisinger, & D. Iliescu (Eds.), *The ITC international handbook of testing and assessment* (pp. 349–363). Oxford, England: Oxford University Press.

Wetzel, E., Böhnke, J. R., Carstensen, C. H., Ziegler, M., & Ostendorf, F. (2013). Do individual response styles matter? Assessing differential item functioning for men and women in the NEO-PI-R. *Journal of Individual Differences, 34*(2), 69-81. doi:10.1027/1614-0001/A000102

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. doi: 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. doi: 10.1177/1073191115583714

Wilkinson, A. E. (1970). Relationship between measures of intellectual functioning and extreme response style. *The Journal of Social Psychology, 81*(2), 271-272. doi: 10.1080/00224545.1970.9922451

Wright, T.A., & Bonnett, D.G. (2007). Job satisfaction and psychological well-being as non-additive predictors of workplace turnover. *Journal of Management, 33*, 141-160. doi: 10.1177/0149206306297582

Wu, P.-C., & Huang, T.-W. (2010). Person heterogeneity of the BDI-II-C and its effects on dimensionality and construct validity: Using mixture item response models. *Measurement and Evaluation in Counseling and Development, 43*, 155-167. doi:10.1177/0748175610384808

Zettler, I., Lang, J. W., Hülsheger, 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. doi: 10.1111/jopy.12172

Zhou, B., & McClendon, M. J. (1999). Cognitive ability and acquiescence. In American Statistical Association (Eds.), *Proceedings of the survey research methods section* (pp. 1003–1012). Alexandria.

Zickar, M., & Gibby, R. E. (2006). A history of faking and socially desirable responding on personality test. In R. L. Griffith & M. H. Petterson (Eds.), *A closer examination of applicants faking behavior* (pp.21-42). Charlotte, NC: IAP-Information Age Publishing.

Zickar, M. J., Gibby, R. E., & Robie, C. (2004). Uncovering Faking Samples in Applicant, Incumbent, and Experimental Data Sets: An Application of Mixed-Model Item Response Theory. *Organizational Research Methods, 7*(2), 168-190. doi: 10.1177/1094428104263674

Ziegler, M., & Kemper, C. J. (2013). Extreme response style and faking: Two sides of the same coin?. In P. Winker, N. Menold, & R. Porst (Eds.), *Interviewers deviations in surveys – impact, reasons, detection and prevention* (pp. 217–233). Frankfurt a.M.: Peter Lang.

Zuckerman & North (1961). Response set and content factors in the California F-scale and the parental attitude research instrument. *Journal of Social Psychology, 53*(3), 199–210. doi: 10.1080/00224545.1961.9922118

2 Using a Mixed IRT Model to Assess the Scale Usage in the Measurement of Job Satisfaction

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.2016.01998

Abstract

This study investigated the adequacy of a rating scale with a large number of response categories that is often used in panel surveys for assessing diverse aspects of job satisfaction. An inappropriate scale usage is indicative of overstraining respondents and of diminished psychometric scale quality. The mixture item response theory (IRT) approach for polytomous data allows exploring heterogeneous patterns of inappropriate scale usage in the form of avoided categories and response styles. In this study, panel data of employees (*n* = 7,036) on five aspects of job satisfaction measured on an 11-point rating scale within the "Household, Income and Labour Dynamics in Australia" survey (wave 2001) were analyzed. A three-class solution of the restricted mixed generalized partial credit model fits the data best. The results showed that none of the three latent classes found used the 11-point rating scale appropriately, and the number of categories used was reduced in all classes. Respondents of the large class (40%) appropriately differentiate between up to six categories. The two smaller classes (33% and 27%) avoid even more categories and show some kind of extreme response style. Furthermore, classes differ in socio-demographic and job-related factors. In conclusion, 2- to 6-point rating scales without the middle point may be more adequate for assessing job satisfaction.

Keywords: job satisfaction, rating scale, large number of response categories, scale usage, response style, mixed IRT models

Using a Mixed IRT Model to Assess the Scale Usage in the Measurement of Job Satisfaction

Job satisfaction is a relevant indicator of quality of life and as such is well investigated in organizational contexts. As Spector (1997, p. vii) pointedly put it, „Job satisfaction is the degree to which people like their jobs". More precisely, the term includes subjective evaluations of relevant work aspects and the affective states the individual is experiencing while at work. Job satisfaction has become important in human resource management, guiding corporation policies in shaping processes and improving effectiveness. High job satisfaction is thought to reflect a good fit of employees' professional and personal characteristics to the job tasks and exhibits a positive effect on commitment and productivity (Judge, Thoresen, Bono, & Patton, 2001). For individuals, high job satisfaction often implies an adequate work-life balance, which in turn increases well-being and life satisfaction (Kossek & Ozeki, 1998).

Because of its importance, a single-item measure of general job satisfaction is often included in national panel surveys, sometimes backed by measures for satisfaction with certain aspects of the job, such as income and relations with colleagues. What is striking is the diversity of response formats across studies. The number of response categories of the rating scale varies considerably: Only four categories were used in the Survey of Health, Ageing and Retirement in Europe (SHARE) as opposed to 11 categories in the German Socio-Economic Panel (GSOEP), the Household, Income and Labour Dynamics in Australia Survey (HILDA), and the Swiss Household Panel (SHP). Such a high number of response categories (a long rating scale) is intended to lead to a measure that reflects the fine-grained differences between subjects in the rated trait (Preston & Colman, 2000). However, it is unclear whether the ratings elicited by a long rating scale can be thought of as representative of the underlying job satisfaction or whether other processes shape the differences found in measure.

2.1 Inappropriate Scale Usage

Answers to a survey questionnaire are based on individuals' knowledge of the topic and their habits of thinking in a certain number of subjectively meaningful categories, for example, black-and-white-thinking versus sophisticated thinking (Naemi, Beal, & Payne, 2009; Viswanathan, Sudman, & Johnson, 2004). Rating scales with very few response categories may not allow for sufficient differentiation, while rating scales with very many categories may overburden individuals (Weng, 2004). Too many as well as too few response categories are therefore a potential source of inappropriate scale usage and bias. Inappropriate scale usage (ISU in the following) refers to individual tendencies in responding unrelated to the content of the question at hand (Paulhus, 1991). In general, simplifying strategies are frequently employed. Pronounced simplifying strategies are commonly known as response styles: the preference for extreme categories (extreme response style, ERS), preference of the middle category (MRS), as well as an acquiescent response style (ARS) and a disacquiescent response style (DARS). Empirically, these response styles are found in major sample portions, for example, for the ERS 25−30% of respondents, for the MRS 11−33%, and for the ARS 32−52% (Carter, Dalal, Lake, Lin, & Zickar, 2011; Meiser & Machunsky, 2008; Wetzel, Carstensen, & Böhnke, 2013). Less pronounced strategies such as avoidance of certain categories have received less attention (for a recent overview, see Van Vaerenbergh & Thomas, 2013; Viswanathan et al., 2004). Eid and Rauber (2000) report that roughly a third of employees in their sample was using only five of the six proposed response categories when asked to rate satisfaction with their superior. If such a misfit between the proposed and the subjectively meaningful number of response categories exists, the rating scale will not adequately reflect the continuous underlying trait and hence violate assumptions for a rating scale (Meiser & Machunsky, 2008).

Empirical results of the effects of rating scale length on response behavior have revealed three important aspects:

1. There is interindividual heterogeneity in scale usage (Jin & Wang, 2014). Studies employing mixed IRT models mostly used data collected with a 4- to 6-point rating scale and often report at least two latent classes of individuals that differ in scale usage, regardless of the proposed rating scale. One latent class often exhibits the ERS. Another one exhibits the MRS when the rating scale included a few response categories (e.g., a 4-point rating scale) but an ordinary scale usage when the number of response categories is increasing up to six options (Eid & Rauber, 2000; Meiser & Machunsky, 2008).
2. Several types of ISU can occur simultaneously. For example, Baumgartner and Steenkamp (2001) reported about high correlations for the ERS with DARS and ARS ($r_{\text{ERS, DARS}} = .41$, $r_{\text{ERS, ARS}} = .59$), and Weijters and colleagues (2010) confirmed these findings ($r_{\text{ERS, DARS}} =.62$, $r_{\text{ERS, ARS}} =.72$). That suggests that within a questionnaire, one may select extreme categories for an item set and disagree or agree with other items regardless of their content. Furthermore, particularly individuals with the ERS are also inclined to reduce the proposed rating scale to a few subjectively meaningful categories (Eid & Rauber, 2000; Meiser & Machunsky, 2008; Wu & Huang, 2010).
3. ISU depends on the trait, population, and context (Kieruj & Moors, 2013).

2.2 Psychometric Quality of Data

High reliability and validity scores are often interpreted to reflect the adequacy of the rating scale (Cox, 1980). For rating scales with two to ten options, the increase of the number of response categories of two to six categories only leads to an increase in reliability and convergent validity measured by a heterotrait-monomethod correlation (Culpepper, 2013; Lozano, García-Cueto, & Muñiz, 2008; Maydeu-Olivares, Kramp, García-Forero, Gallardo-Pujol, & Coffman, 2009). However, ISU is responsible for up to 25% of score variability (Wetzel & Carstensen, 2017) and can thereby make a contribution to the artificial increase of reliability (Jin & Wang, 2014; Weathers, Sharma, & Niedrich, 2005). A separate assessment of true trait variance and response style variance is necessary to obtain an unbiased reliability score. Chang (1994) demonstrated that a 6-point rating scale contains a higher proportion of response style variance compared to a 4-point rating scale, thus reducing psychometric quality. Weather and colleagues (2005) explained that particularly individuals with limited cognitive resources (e.g., low discrimination ability) react with intensive ISU when confronted with a long rating scale. Another important finding is that the reliability of homogeneous scales is less affected by rating scale length (Weng, 2004). In general, a response format with four to six or seven response categories is considered optimal with regard to psychometric quality (Chang, 1994; Culpepper, 2013; Lozano et al., 2008; Weng, 2004). However, in large-scale panel studies, an 11-point rating scale and a few items are widely and considered as the gold standard of satisfaction assessment. To our knowledge, it has not been analyzed whether this rating scale is appropriate or produces ISU. Given the results of previous studies, individuals may differ in response style use. Some individuals may be overwhelmed by 11 response categories, whereas others might not have any problem with such a long rating scale.

2.3 The Mixture Item Response Theory Approach for Polytomous Data

A suitable alternative for examining the adequacy of an 11-point rating scale is the mixture IRT approach, which allows modeling the response process on the level of single items and categories and detecting heterogeneous scale usage. Mixed IRT models such as the mixed partial credit model (mPCM; Rost, 1997) can be applied to investigate a number of scale characteristics. The focus can lie on (a) heterogeneity of scale usage (e.g., Eid & Rauber, 2000); (b) appropriate usage of particular response categories (e.g., Carter et al., 2011); (c) an adequate representation of the continuity of a trait by the rating scale (e.g., Meiser & Machunsky, 2008), or on (d) stability of scale usage across items, subscales or different traits (e.g., Wetzel et al., 2013).

The number and size of latent classes are unknown and result by applying a mixed IRT model. The qualitative differences between classes in scale usage are detectable from the interpretation of class-specific item parameters and item profiles (Rost, 1997). Different types of ISU can be distinguished (e. g., “avoided” or unused categories, response styles, or socially desirable responding; see Eid & Zickar, 2007; Wetzel et al., 2013; Wu & Huang, 2010). Further, the resulting class-specific latent trait values of individuals can be estimated. In contrast to raw total scores, those are adjusted to the effect of class-specific scale usage and can be used to accurately compare individuals within and across latent classes (Eid & Rauber, 2000).

While the mPCM has been widely applied in the area of personality scales (e.g., Eid & Zickar, 2007; Maij-de Meij, Kelderman, & van der Flier, 2008; Meiser & Machunsky, 2008; Wetzel et al., 2013), systematic research on how job satisfaction can be measured appropriately using the mixture IRT approach is still lacking. In the current paper, we will focus on evaluating the appropriateness of the long rating scale (including 11 response categories) that is used in the HILDA survey for assessing different aspects of job satisfaction. Because we assume that items will differ in their discrimination power, we will apply a mixture distribution IRT model with varying item discrimination parameters to test whether there are classes that use this rating scale in different ways.

2.4 Materials and Methods

2.4.1 Sample

We used data from the first wave (collected in 2001) of the HILDA survey. The HILDA survey is Australia's nationally representative household-based panel study.[6] The data collection is primarily focused on subjective well-being, income and welfare, family formation, and labor market dynamics. The survey is conducted by the Melbourne Institute of Applied Economic and Social Research, from which the license for the data set can be obtained (Summerfield, Freidin, Hahn, Li, Macalalad, Mundy, et al., 2015). While the general sample of the first wave consists of 13,969 individuals, a subsample of 7,036 subjects was obtained by the following inclusion criteria: A minimum age of 18, paid employment, and no missing values for the items on job satisfaction[7]. This sample consists of about half women (47.1 %). The overall mean age is 39.2 years (*SD* = 11.48, *max* = 82). Concerning the level of education, 57.6 % have at least a graduate degree. Most subjects are employees (93%) and most work full time (73.7%).

2.4.2 Measures

Job Satisfaction (JS)

The HILDA survey includes 5 items on satisfaction with various aspects of the current job: total pay, job security, work itself, working hours, and flexibility to balance work and non-work commitments. An 11-

[6] This paper uses unit record data from the Household, Income and Labour Dynamics in Australia (HILDA) Survey. The HILDA Project was initiated and is funded by the Australian Government, Department of Social Services (DSS) and is managed by the Melbourne Institute of Applied Economic and Social Research (Melbourne Institute). The findings and views reported in this paper, however, are those of the authors and should not be attributed to either the DSS or the Melbourne Institute.

[7] Proportion of cases with missing values on these items ranged from 0.1−0.3%.

point rating scale (0 = *totally dissatisfied* and 10 = *totally satisfied*) was used as a response format.

Predictor variables

Job position (JP) is a single-item measure. The eight categories of the original variable were regrouped into the following hierarchical work levels: Specialists and executive staff (level 1), administrators (level 2), staff of the service sector such as machinery operators, drivers, and so forth (level 3).

Organization size (OS) is derived from an item on the number of persons employed in the organization of respondents. We defined an organization with fewer than 20 persons as small, with 20−200 persons as medium-sized, and with more than 200 persons as a large one.

Job Characteristics (JC) were measured with 10 items. Respondents evaluated their psychosocial work conditions using a 7-point rating scale from 1 (*strongly disagree*) to 7 (*strongly agree*). Based on an exploratory factor analysis, we reduced the 10 items to four aspects of working conditions: (1) *stress* (e.g., "My job is more stressful than I had ever imagined."), (2) *security* (e.g., "The company I work for will still be in business 5 years from now."), (3) *autonomy* (e.g., "I have a lot of say about what happens on my job."), and (4) *skills* (e.g., "I use many of my skills and abilities in my current job.") See the appendix to Chapter 2 for more details.

The importance of employment and work situation is measured with one item using an 11-point rating scale from 0 (*the least important thing*) to 10 (*the most important thing*). Further single-item measures were ***tenure in respondents' current occupation*** (in years), and ***total financial year income*** (AUD$ in thousands).

2.4.3 Statistical Analyses

2.4.3.1 Non-Technical Introduction into the IRT models

First, we will present a general unidimensional logistic model for ordered categorical data, the generalized partial credit model (GPCM; Muraki, 1992). We then describe a more general model incorporating latent classes, the mixed distribution generalized partial credit model (mGPCM; Von Davier & Yamamoto, 2004), and its more parsimonious variant, the restrictive mixed generalized partial credit model (rmGPCM).

The Generalized Partial Credit Model

The GPCM (Muraki, 1992) extends the partial credit model (PCM; Masters, 1982) to include item-specific discrimination parameters. Thus, the GPCM contains two types of parameters that link the manifest item responses to the continuous latent continuum: Item-specific threshold parameters τ_{is} that locate the "skip" between two adjacent categories, $x - 1$ and x, and item-specific discrimination parameters δ_i that characterize the discrimination power of an item. The response probability $P_{vix}(\theta)$ for category x ($x \in \{0, \dots, m\}$) of a polytomous item i given the latent trait score θ_v of an individual v is modeled using these two types of item parameters:

$$P_{vix}(\theta) = \frac{\exp[\sum_{s=0}^{x} \delta_i(\theta_v - \tau_{is})\,]}{\sum_{c=0}^{m} \exp[\sum_{s=0}^{c} \delta_i(\theta_v - \tau_{is})]} \tag{2.1}$$

with $\delta_i > 0$, $E(\theta_v) = 0$, and $\sum_{s=0}^{0} \delta_i(\theta_v - \tau_{is}) = 0$ for all i.

These response probabilities can be depicted in the form of category characteristic curves (CCCs). Figure *2.1* shows the CCCs of two fictitious items with 11 response categories ($x = 0, \dots, 10$). These items have identical threshold parameters and differ only in their discrimination parameter. For both items, the response probability for the first category is monotonically decreasing. With increasing latent trait value, selecting the first category becomes less likely. In an analogue manner, the probability for the last category is monotonically increasing, indicating that selecting this category becomes more likely with increasing θ value. The CCCs of the remaining categories are unimodal. The intersections of CCCs of adjacent categories $x - 1$ and x are represented by threshold parameters τ_{is}. In general, there are $m - 1$ threshold parameters for each item. Thresholds are placed on the same latent continuum θ. The lower the threshold, the easier it is to choose the higher of two adjacent categories given the latent trait value θ_v. The additional item-specific discrimination parameters δ_i that distinguish the GPCM from the PCM also affect response probabilities $P_{vix}(\theta)$ in a particular category x of item i at the latent trait values θ_v. The higher the discrimination parameter, the steeper and narrower are the CCCs. In Figure *2.1*, the dashed lines represent the CCCs of an item with a higher discrimination parameter compared to another item which CCCs are depicted in solid lines.

The Mixed Generalized Partial Credit Model

The mGPCM (Von Davier & Yamamoto, 2004) is an extension of the GPCM and assumes the existence of a priori unobserved subpopulations. The mGPCM is defined by the following equation:

$$P_{vix}(\theta) = \sum_{g=1}^{G} \pi_g \frac{\exp[\sum_{s=0}^{x} \delta_{ig}(\theta_{vg} - \tau_{isg})]}{\sum_{c=0}^{m} \exp[\sum_{s=0}^{c} \delta_{ig}(\theta_{vg} - \tau_{isg})]} \tag{2.2}$$

with $\sum_{g=1}^{G} \pi_g = 1$, $E(\theta_{vg}) = 0$ for all g, $\sum_{s=0}^{0} \delta_{ig}(\theta_{vg} - \tau_{isg}) = 0$ for all i in all g.

Each parameter of the mGPCM obtains an additional index *g* (*g* ∈{1,..., *G*}), which indicates a latent class. π_g $(0 < \pi_g < 1)$ is the size of latent class *g*. The number of latent classes *G* is no a model parameter but is determined by comparing models with a different number of classes by means of goodness-of-fit statistics. In the mGPCM, threshold and discrimination parameters are class-specific. Therefore, the CCCs differ between latent classes and can be used to identify peculiar scale usage patterns within a homogeneous class.

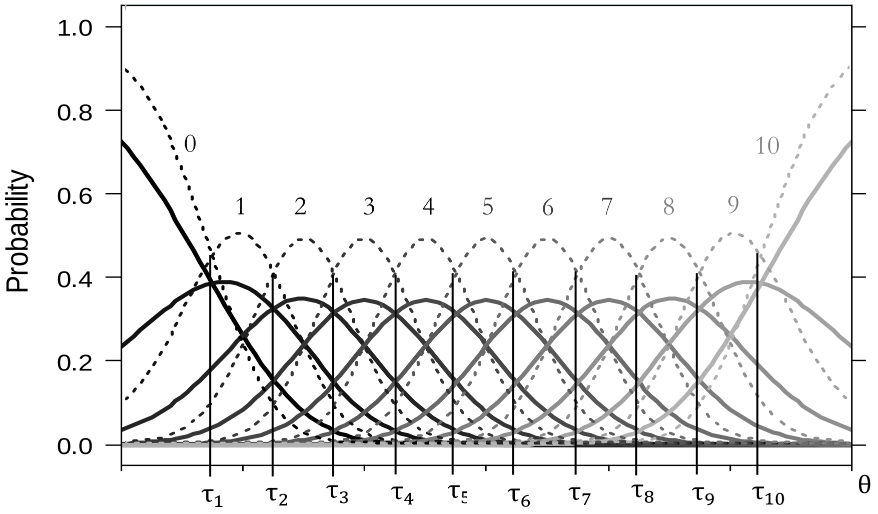


Figure 2.1. Category characteristic curves for two fictitious items measured with an 11-point rating scale (solid lines for item 1a and dashed lines for item 1b). Two items share the same ordered threshold parameters (τ_1 = - 2.25, τ_2 = -1.75, τ_3 = - 1.25, τ_4 = -0.75, τ_5 = -0.25, τ_6 = 0.25, τ_7 = 0.75, τ_8 =1.25, τ_9 = 1.75, τ_{10} = 2.25) and differ in their discrimination parameters (δ_a = 1.50, δ_b = 3.00).

Restricted model version

Compared to the mGPCM, the rmGPCM assumes equal discrimination parameters across latent classes (but not items):

$$P_{vix}(\theta) = \sum_{g=1}^{G} \pi_g \frac{\exp\left[\sum_{s=0}^{x} \delta_i\left(\theta_{vg} - \tau_{isg}\right)\right]}{\sum_{c=0}^{m} \exp\left[\sum_{s=0}^{c} \delta_i\left(\theta_{vg} - \tau_{isg}\right)\right]} \tag{2.3}$$

with $\sum_{g=1}^{G} \pi_g = 1$, $E(\theta_{vg}) = 0$ for all *g*, $\sum_{s=0}^{0} \delta_i(\theta_{vg} - \tau_{isg}) = 0$ for all *i* in all *g*.

By contrast with the mGPCM, discrimination parameters δ_i lack index *g*. Moreover, in the mPCM these discrimination parameters are constrained to be equal across items (for details concerning the mPCM, see Carter et al., 2011).

In the present study, we will start with an application of the more parsimonious rmGPCM to the five JS items, because we suspect that the mGPCM (with item-specific discrimination parameters within latent classes) is too complex to fit well. We first determine the best model solution of the rmGPCM. In the next step, we compare the best solution of the rmGPCM to the mPCM, which is more restrictive, and the mGPCM, which is more general. These model comparisons would reveal whether including discrimination parameters in a model improves the model-data fit. Finally, we explain the assignment of individuals to latent classes found in the best-fitted model solution using socio-demographic and job-related variables.

2.4.3.2 Estimation

For estimating the rmGPCM (and above-mentioned model variants), the Latent GOLD 5.0 software package was used (Vermunt & Magidson, 2013). Here, the marginal maximum likelihood function (MML) is maximized using an EM algorithm initially, switching to the Newton-Raphson method in the end. The number of iterations was set to 8,000 and 600, respectively, and 100 sets of starting values were used (see the appendix to Chapter 2 for the syntax).

2.4.3.3 Model Fit

In the first step, the adequate number of classes was determined by comparing rmGPCMs with one to five classes with regard to the consistent Akaike information criterion (CAIC; Bozdogan, 1987), which is suitable for comparison of mixture IRT models with a varying number of subpopulations (Cho, 2013). The class solution with the lowest CAIC value indicates the preferable model. Additionally, in order to test whether the expected frequencies of response patterns in the selected model deviated significantly from the observed ones in the empirical data, we calculated parametric bootstrapping p-values for the Pearson and Cressie-Read $\chi2$ goodness-of-fit statistics using 500 bootstrapping samples (default). In a second step, we tested whether model fit was improved by estimating (a) the more parsimonious mPCM, which assumes equal discrimination parameters across items, or (b) the more general mGPCM, which assumes item-specific discrimination parameters within classes. These models were compared to the rmGPCM by conducting bootstrapping $\chi2$-difference tests. A significant test result indicates a better fit of the more complex model.

2.4.3.4 Exploring Scale Usage

Mapping the item parameters and CCCs to scale usage of the best-fit model, the following three aspects are important: (1) Ordered thresholds mean that all proposed response categories are present on the latent continuum in ascending order. Figure *2.1* represents the ordinary scale usage of two items in which the thresholds are ordered and equidistant. Unordered thresholds often indicate avoided categories (Wetzel & Carstensen, 2014). In Figure *2.2*, for two items, the order of τ_{i8} and τ_{i7} is reversed, indicating a complete overlap of the CCC pertaining to category x = 7 by the CCCs of categories x = 6 and x = 8. Apparently, this category is avoided. While a class may exhibit ordered thresholds, another one may be characterized by the omission of certain categories. (2) The distance between adjacent thresholds represents the width of a particular category. The wider is the distance between thresholds, the larger is the segment of the latent continuum represented by this category. Classes may also differ in response styles, with, for example, an ERS class characterized by very wide extreme latent categories. (3) Higher discrimination parameters lead to less overlap between CCCs. Also, response probabilities then change more rapidly with increasing latent trait value.

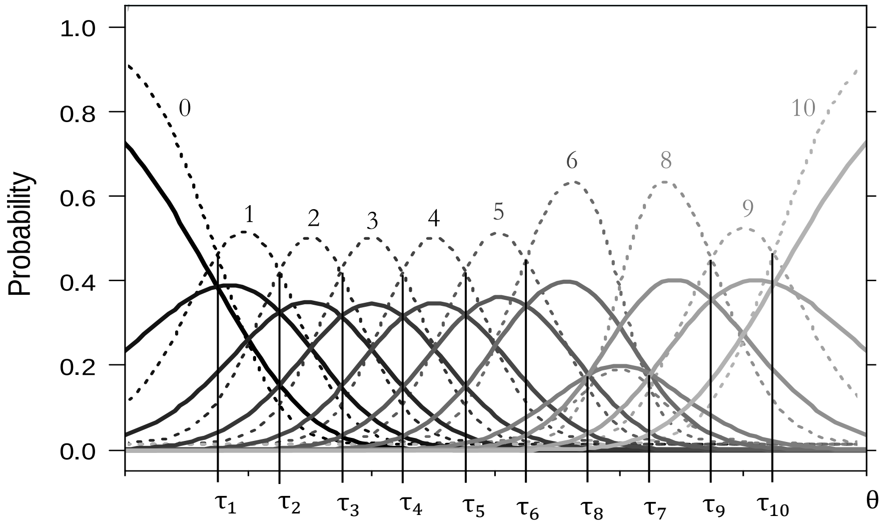


Figure 2.2. Category characteristic curves for two fictitious items measured with an 11-point rating scale (solid lines for item 2a and dashed lines for item 2b). Two items share the same partly disordered threshold parameters (τ_1= - 2.25, τ_2= -1.75, τ_3= -1.25, τ_4= -0.75, τ_5= -0.25, τ_6= 0.25, τ_7= 1.25, τ_8= 0.75, τ_9 = 1.75, τ_{10} = 2.25), but differ in their discrimination parameters (δ_a = 1.50, δ_b = 3.00).

2.4.3.5 Predicting Class Assignment

Because the JC subscale scores showed a non-ignorable amount of missing data (8.5–9.4%) multiple imputation was applied (Enders, 2010). Following recommendations by Graham and colleagues (2007), we generated 20 data sets with missing values on the JC subscales replaced by the means of a sequential regression method (as implemented in IBM SPSS Statistics package v23, IBM Corp., Armonk, NY, USA). The imputation model included all predictor variables, latent class assignment, and estimated person parameters in JS obtained from the best class-solution of the mixed IRT model, as well as personality traits such as conscientiousness that predicted the missingness (see the appendix to Chapter 2 for details on the missing analysis). The following analysis was automatically performed on the 20 generated data sets and results were subsequently aggregated. Assignment to latent classes was predicted in a multinomial logistic regression model. Classification inaccuracy of the rmGPCM was taken into account by using the adjusted three-step method proposed by Vermunt (2010). That method is implemented in Latent GOLD 5.0. For categorical predictors (e.g., job position, organization size), sets of dummy variables were built. To reduce the number of dummy variables, the categories of original predictor variables were regrouped as described above.

2.5 Results

2.5.1 Descriptive Analysis

Table 2.1 provides descriptive statistics for the JS items. The relative frequencies demonstrate that response categories in the lower part of the response format are underrepresented. In particular, the two lowest categories were chosen by less than 2% of the sample. The category chosen most frequently was either category $x = 8$ (for aspects concerning *total pay*, *work itself*, and *working hours*) or category $x = 10$ (for aspects concerning *job security* and *flexibility*).

Table 2.1. Descriptive statistics for aspects of job satisfaction.

	Statistics		**Relative Category Frequencies**										
Item label	***M***	***SD***	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Total pay	6.73	2.41	1.9	1.6	3.2	4.9	5.1	10.8	10.0	18.5	21.4	9.7	12.8
Job security	7.72	2.50	1.9	1.5	2.5	2.9	2.9	7.1	4.6	10.0	18.6	17.6	30.8
Work itself	7.67	2.06	0.5	0.7	1.4	2.4	2.6	7.3	7.7	15.2	23.6	17.5	21.1
Working hours	7.14	2.35	1.0	1.4	2.7	3.9	4.3	10.8	8.7	15.1	20.8	13.3	18.0
Flexibility	7.36	2.60	2.0	1.7	3.2	3.8	3.8	8.5	6.3	11.6	18.0	15.0	26.0

2.5.2 Determination of the Number of Latent Classes

All estimated models reached convergence. It took between 2 and 252 iterations in EM algorithm and 2 to 11 iterations in the Newton-Raphson algorithm. For the rmGPCM with one to five classes, goodness-of-fit statistics can be found in Table 2.2. The three-class rmGPCM indicated the best relative model fit (the lowest $\text{CAIC}_{\text{rmGPCM-3}}$ = 137637). Also, with respect to absolute fit, the three-class rmGPCM fitted the data very well (p > .05 for bootstrapped Pearson and Cressie-Read χ^2 statistics).

The three-class rmGPCM was then compared to the three-class mPCM and the three-class mGPCM. The three-class rmGPCM shows a better fit to the data than the mPCM (lower CAIC, $\Delta\chi^2(4)$ = 272.81; bootstrapped p < .001). The rmGPCM and mGPCM demonstrated no statistically relevant differences in their data fit (slight difference in the CAIC values; $\Delta\chi^2(10)$ = 51.87, bootstrapped p = .94). Thus, we accepted the more parsimonious three-class rmGPCM. More details for class solutions of the mPCM, rmGPCM, and mGPCM are provided in Table 2.7 in the appendix to Chapter 2.

Table 2.2. Goodness-of-fit statistics for the rmGPCM and competing models.

Model	N_{par}	**LL**	**CAIC**	**Pearson *p*-Value**	**CR *p*-Value**	**BV**	**Extr.** τ_{isg}	**Extr. SE**	**Bootstrapped** χ^2**-Difference Test (*df*) *p*-Value**
rmGPCM									
1 class	55	-69917	140376						
2 classes	107	-68382	137819						
3 classes	159	-68035	**137637**	.99	.99	0	1	7	
4 classes	211	-67877	137833						
5 classes	263	-67736	138065						
mPCM 3 classes	155	-68173	137873			0	2	5	rmGPCM vs. mPCM (3 cl): 272.81 (4) < .001
mGPCM 3 classes	169	-68009	137684			2	98	7	mGPCM vs. rmGPCM (3 cl): 51.87 (10) p = .94

Notes. N_{par}: the number of model parameter. LL: Log-Likelihood. CAIC: Consistent Akaike's Information Criterion. Pearson p-Value: the bootstrapped p-value corresponding to the Pearson χ^2 goodness-of-fit statistic. CR p-Value: the bootstrapped p-value corresponding to the Cressie-Read χ^2 goodness-of-fit statistic. BV: boundary values. Extr. τ_{isg}: the number of threshold parameters larger than $|4|$. Extr. SE: the number of extreme standard errors of item parameters. (Extreme standard errors are defined as values five times larger than the most frequently occurring standard errors in the estimated model. Here, larger than 1.5). The lowest CAIC is marked in boldface.

2.5.3 Class-Specific Scale Usage

Based on the rmGPCM-3, individuals were first assigned to latent classes by using their largest class assignment probability value. To evaluate the accuracy of classification, the mean assignment probability for each latent class was calculated. It can be considered as good and equals to .79 for the first class (as ordered by size, $\pi_1 = .40$), to .85 for the second class ($\pi_2 = .33$), and to .76 for the third class ($\pi_3 = .27$). Table 2.3 presents the class-specific item parameters and the corresponding robust standard errors of the three-class rmGPCM. The category characteristic curves are shown in Figure 2.3. The classes differ with regard to scale usage in the following way: While in the first (largest) class, at least half of the thresholds (five to eight, depending on a specific item) are in the order expected given the response format, this only holds true for two or three thresholds in the second class. The third (smallest) class can be placed between these two classes, with four to six correctly ordered thresholds. This implies major deviations of class-specific response patterns from the ordinary scale usage in all classes. More evidence for presumed class-specific scale usage can be drawn from the distances between adjacent thresholds, which are far from equidistant ($\text{range}_{\text{class 1}}$ [0.28, 1.27], $\text{range}_{\text{class 2}}$ [0.67, 2.00], $\text{range}_{\text{class 3}}$ [0.37, 1.62]). Taken together, the most refined differentiation between response categories can be expected in the first class, a moderate one in the third class, and the crudest in the second class. We will now look more closely at the scale usage within each class.

2.5.3.1 Class 1

In this largest class, the first four thresholds are mainly ordered and rather equidistant. For the items *work itself* and *working hours*, this pattern is slightly altered because of extreme parameter estimates, most likely due to the low category frequencies. However, one can infer that in this class, individuals with a low latent trait level are expected to differentiate between the lower categories ($x = 0$ to $x = 3$). In the medium segment of the latent continuum, thresholds are mostly unordered, and only categories $x = 5$ and $x = 7$, if at all, appear. Remarkable in the upper segment of the latent continuum is the reversed order of the 9th and 10th thresholds, so that category $x = 9$ is completely covered by the wide neighboring categories, suggesting that in the first class this category is generally avoided. Taken together, there is refined differentiation in the rating of low job satisfaction in this class, while above average and highly satisfied workers seem to decide only between two categories. Of the 11 response categories proposed in the manifest rating scale, only five to six are represented on the latent continuum. Based on the particularities described above, we label the scale usage of this class *differential* response style (DRS).

Table 2.3. Latent class-specific item parameters of the three-class rmGPCM.

Item label	δ_i	τ_{i1g}	τ_{i2g}	τ_{i3g}	τ_{i4g}	τ_{i5g}	τ_{i6g}	τ_{i7g}	τ_{i8g}	τ_{i9g}	τ_{i10g}
	Class 1 ($\pi_1 = .40$)										
Total Pay	1 (-)	-2.27 (1.91)	-1.52 (0.44)	-0.86 (0.22)	-0.41 (0.14)	-0.64 (0.12)	-0.25 (0.10)	-0.46 (0.08)	-0.10 (0.08)	1.92 (0.26)	1.16 (0.38)
Job security	0.71 (0.04)	-2.43 (1.63)	-1.72 (0.47)	-1.18 (0.24)	0.04 (0.16)	-1.16 (0.15)	-0.06 (0.13)	-0.84 (0.11)	-0.99 (0.09)	0.95 (0.14)	0.29 (0.14)
Work itself	1.27 (0.08)	2.44 (8.30)	-5.71 (6.66)	-1.42 (0.57)	-0.45 (0.23)	-0.83 (0.16)	-0.30 (0.12)0	-0.53 (0.09)	-0.36 (0.07)	1.08 (0.19)	0.95 (0.30)
Working hours	2.58 (0.24)	-3.47 (2.21)	-1.28 (2.02)	-0.66 (0.33)	-0.41 (0.19)	-0.41 (0.14)	-0.11 (0.11	-0.23 (0.09)	-0.03 (0.09)	0.86 (0.36)	0.67 (0.49)
Flexibility	1.76 (0.15)	-1.08 (0.96)	-1.06 (0.53)	-0.56 (0.23)	-0.22 (0.17)	-0.48 (0.16)	-0.17 (0.12)	-0.33 (0.09)	-0.16 (0.08)	0.56 (0.13)	0.53 (0.22)
	Class 2 ($\pi_2 = .33$)										
Total Pay	1 (-)	0.98 (0.29)	-0.77 (0.31)	-0.52 (0.22)	-0.08 (0.20)	-1.22 (0.16)	0.61 (0.15)	-0.78 (0.14)	-0.24 (0.10)	1.24 (0.18)	-1.89 (0.18)
Job security	0.71 (0.04)	1.11 (0.29)	-1.00 (0.30)	0.10 (0.26)	-0.05 (0.27)	-1.79 (0.27)	1.47 (0.21)	-0.80 (0.24)	-1.35 (0.18)	0.98 (0.20)	-3.51 (0.19)
Work itself	1.27 (0.08)	0.22 (0.45)	-0.84 (0.47)	-0.44 (0.32)	-0.06 (0.32)	-1.39 (0.27)	0.25 (0.17)	-0.64 (0.16)	-0.49 (0.12)	0.74 (0.18)	-1.61 (0.17)
Working hours	2.58 (0.24)	0.12 (0.48)	-0.77 (0.48)	-0.26 (0.27)	-0.15 (0.28)	-0.77 (0.23)	0.21 (0.17)	-0.34 (0.18)	-0.29 (0.13)	0.56 (0.24)	-0.88 (0.23)
Flexibility	1.76 (0.15)	0.30 (0.30)	-0.57 (0.33)	-0.18 (0.27)	-0.16 (0.28)	-0.93 (0.23)	0.63 (0.23)	-0.69 (0.23)	-0.38 (0.14)	0.54 (0.19)	-1.38 (0.17)
	Class 3 ($\pi_3 = .27$)										

Item label	$\boldsymbol{\delta_i}$	$\boldsymbol{\tau_{i1g}}$	$\boldsymbol{\tau_{i2g}}$	$\boldsymbol{\tau_{i3g}}$	$\boldsymbol{\tau_{i4g}}$	$\boldsymbol{\tau_{i5g}}$	$\boldsymbol{\tau_{i6g}}$	$\boldsymbol{\tau_{i7g}}$	$\boldsymbol{\tau_{i8g}}$	$\boldsymbol{\tau_{i9g}}$	$\boldsymbol{\tau_{i10g}}$
Total Pay	1 (-)	-1.61 (0.50)	-0.60 (0.26)	-0.35 (0.22)	0.22 (0.25)	-0.83 (0.23)	-0.01 (0.19)	-0.85 (0.16)	-0.03 (0.13)	-0.11 (0.18)	1.93 (0.25)
Job security	0.71 (0.04)	-0.38 (0.35)	-0.28 (0.32)	0.61 (0.41)	-0.68 (0.41)	-1.43 (0.27)	1.33 (0.35)	-2.24 (0.35)	-0.33 (0.22)	-1.35 (0.22)	0.93 (0.14)
Work itself	1.27 (0.08)	-1.79 (0.74)	-0.47 (0.36)	-0.27 (0.37)	0.01 (0.43)	-0.84 (0.38)	-0.26 (0.26)	-0.68 (0.21)	-0.16 (0.17)	-0.76 (0.21)	1.15 (0.20)
Working hours	2.58 (0.24)	-2.15 (3.82)	-0.25 (0.32)	-0.31 (0.36)	-0.09 (0.38)	-0.59 (0.38)	-0.08 (0.22)	-0.24 (0.20)	-0.05 (0.19)	-0.30 (0.22)	0.79 (0.26)
Flexibility	1.76 (0.15)	-0.66 (0.36)	-0.39 (0.30)	-0.15 (0.33)	-0.18 (0.31)	-0.59 (0.30)	0.18 (0.24)	-0.36 (0.24)	-0.31 (0.19)	-0.45 (0.19)	0.40 (0.15)

Notes. Threshold parameters τ_{isg} are transformed from differences between two adjacent categories parameters $\beta_{0ixg} - \beta_{0ix-1g}$ obtained in Latent GOLD Regression submodule, as follows $\tau_{isg} = -1 * (\beta_{0ixg} - \beta_{0ix-1g})/\ \delta_i$ (Vermunt & Magidson, 2006). Robust standard errors in brackets are calculated by Latent GOLD for the parameters $\beta_{0ixg} - \beta_{0ix-1g}$.

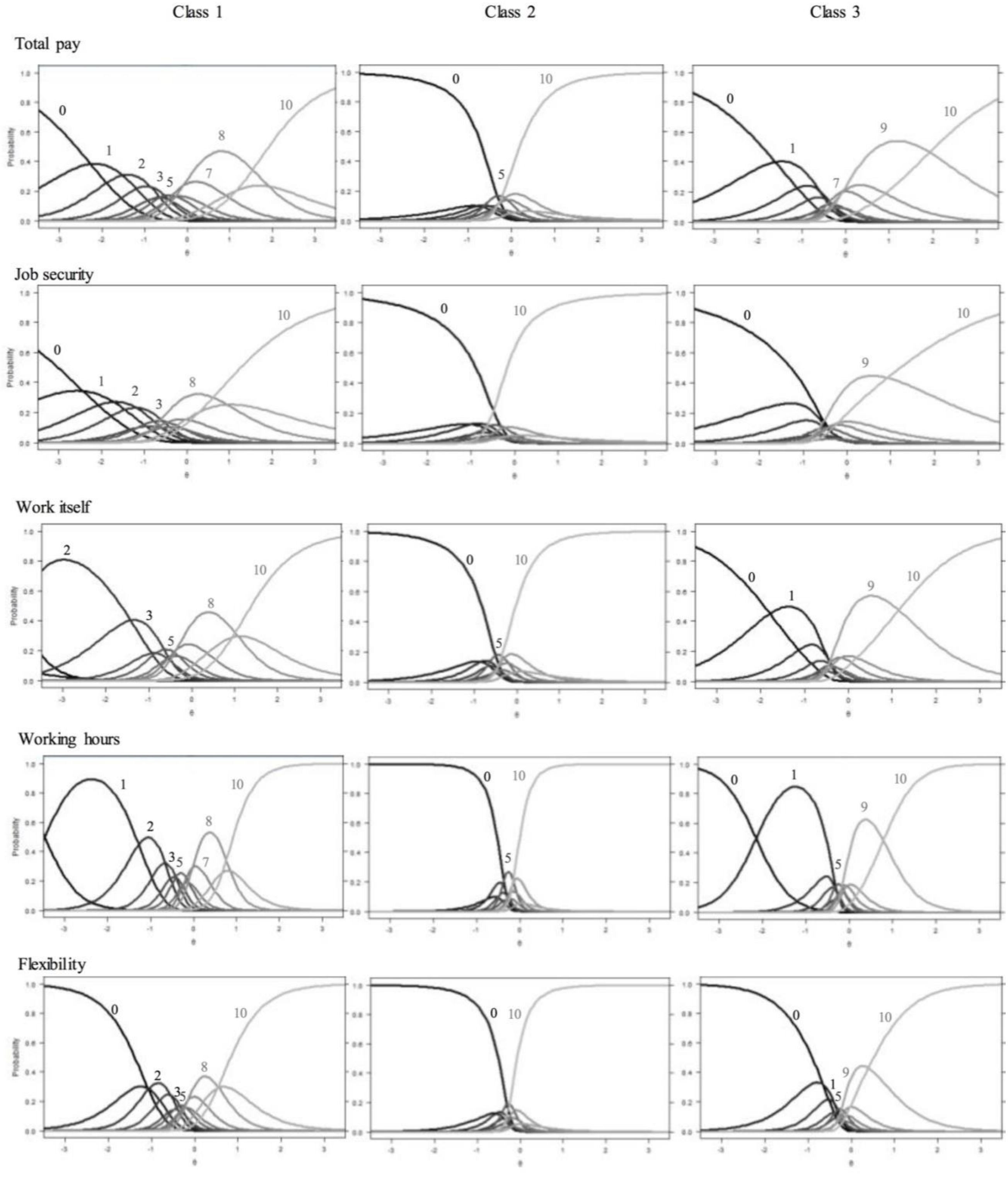


Figure 2.3. Category characteristic curves for aspects of job satisfaction in the three latent classes.

(*Note*. A number indicates the value of a category which has the highest response probability on a certain segment of the latent continuum.)

2.5.3.2 Class 2

The medium-sized class exhibits the most unordered thresholds and the CCCs show the same pattern for all items: The two extreme categories ($x = 0$ and $x = 10$) dominate with high probabilities for a wide segment of the latent trait levels. For items *job security* and *flexibility*, these categories even intersect, indicating a dichotomous response pattern. For the remaining three items, there is a narrow medium segment for which category $x = 5$ is most likely to be chosen. Because of this predominance of the two extreme categories, we label the scale usage of this second class *extreme* response style (ERS).

2.5.3.3 Class 3

For the small class, three to five categories appear to represent the full range of the latent continuum. For items *job security* and *flexibility*, the lowest category covers most of the lower half of the latent continuum, while for the remaining items, there is also a considerable area in which the second category ($x = 1$) is most likely. In the upper half of the latent continuum, the two highest categories ($x = 9$, $x = 10$) of all items dominate all remaining ones. In addition, there is a minimal segment for the middle category for items *total pay* ($x = 7$) and *working hours* ($x = 5$). Taken together, the latent continuum is mostly reduced to four segments which are almost the same width. Because the dominating categories are the extreme ones, we label the scale usage of the third class *semi-extreme* response style (semi-ERS).

2.5.4 Expected Category Frequencies for Job Satisfaction Items in Latent Classes

While evaluation of CCCs allows identifying scale usage patterns, the distribution of the latent variable (job satisfaction) may also differ between latent classes. The expected category frequencies (see Figure 2.4) reflect both, differences in item parameters as well as differences in the distribution of latent variable between classes. Because all classes exhibit low expected frequencies in the first five categories, the sample is quite satisfied on average. Differences between classes emerge in the upper categories. In the distribution of the first (DRS) class, the preference for category $x = 8$ becomes apparent, while for the ERS class, category $x = 10$ resp. category $x = 9$ for the semi-ERS class are expected to be most frequent.

2.5.5 Predicting Class Assignment

Results of the multinomial logistic regression for multiple imputed data sets can be found in Table 2.4. The assignment to the ERS class compared to the DRS class is more likely for female and part-time employees, and those with higher perceived job skills, job security, and greater importance of the job.

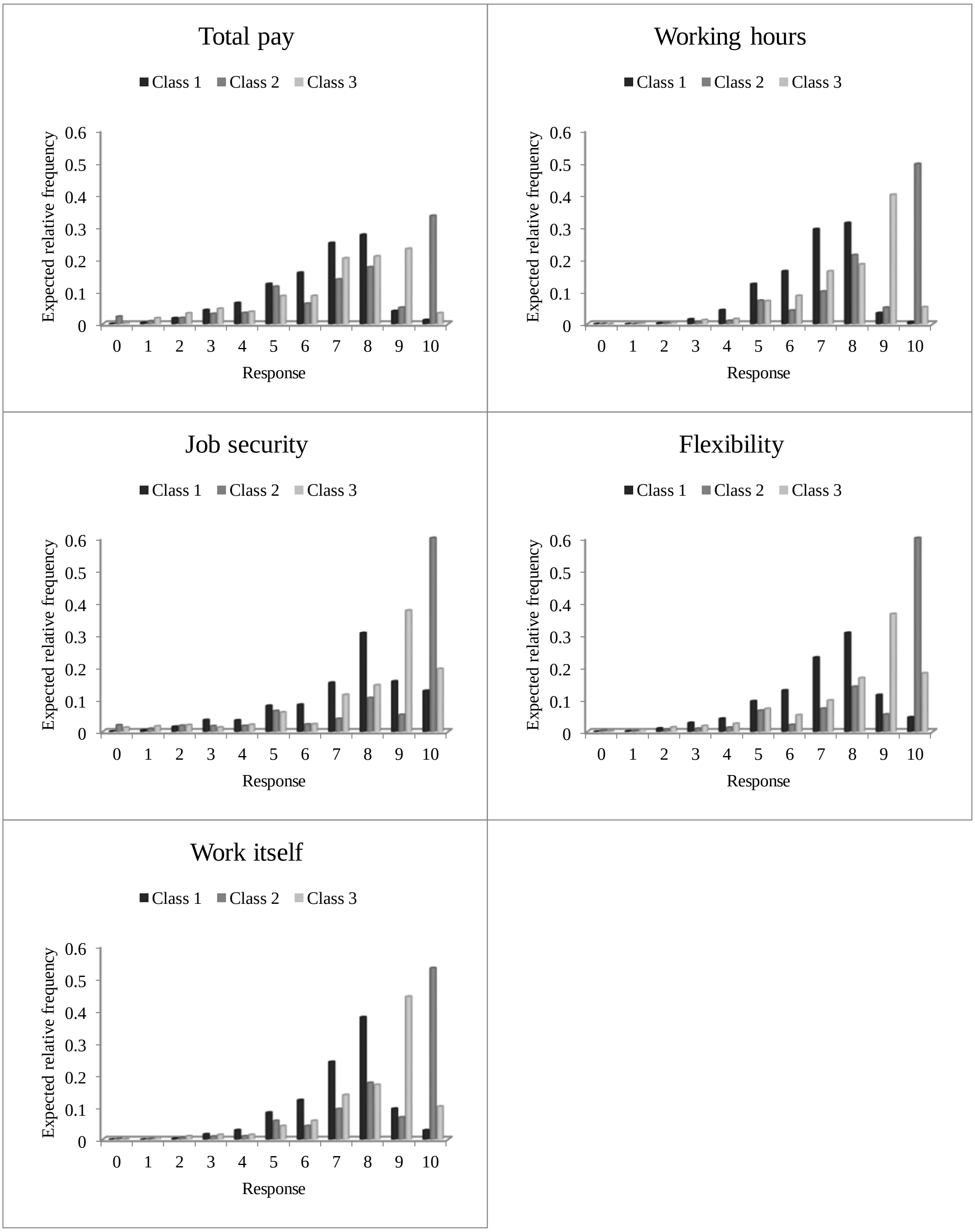


Figure 2.4. Relative frequencies for the eleven response categories of the job satisfaction items expected on the basis of the rmGPCM-3 in the three latent classes.

Table 2.4. Prediction of latent class assignment by means of multinomial regression model.

	ERS class versus DRS class		**Semi-ERS class versus DRS class**		**ERS class versus Semi-ERS class**	
	B **(SE)**	***e*^{b} [95% CI]**	***B*** **(SE)**	***e*^{b} [95% CI]**	***B*** **(SE)**	***e*^{b} [95% CI]**
Constant	-7.16*** (0.49)		-4.53*** (0.52)		-2.63*** (0.53)	
Age	0.02*** (0.00)	1.02 [1.01; 1.03]	0.01 (0.01)	1.01 [1.00; 1.02]	0.01** (0.01)	1.01 [1.00; 1.02]
Gender (female)	0.42*** (0.10)	1.53 [1.25; 1.86]	0.36** (0.12)	1.44 [1.14; 1.81]	0.06 (0.11)	1.06 [0.85; 1.32]
Education level (> 12 years)	-0.30** (0.10)	0.74 [0.61; 0.90]	0.02 (0.12)	1.02 [0.81; 1.28]	-0.32** (0.11)	0.73 [0.59; 0.90]
Income	-0.00 (0.00)	1.00 [1.00; 1.00]	0.00 (0.00)	1.00 [1.00; 1.01]	-0.00 (0.00)	1.00 [0.99; 1.00]
Tenure in current position	0.00 (0.01)	1.00 [0.99; 1.01]	0.00 (0.01)	1.00 [0.99; 1.02]	-0.00 (0.01)	1.00 [0.99; 1.01]
Job position (Level 2)	-0.20 (0.13)	0.82 [0.63; 1.06]	-0.07 (0.16)	0.93 [0.68; 1.27]	-0.13 (0.14)	0.88 [0.67; 1.16]
Job position (Level 1)	-0.43*** (0.13)	0.65 [0.51; 0.83]	-0.27 (0.15)	0.76 [0.57; 1.03]	-0.16 (0.14)	0.85 [0.65; 1.11]
Part-time occupation	0.91*** (0.12)	2.49 [1.96; 3.16]	0.57*** (0.14)	1.76 [1.33; 2.33]	0.35** (0.12)	1.41 [1.12; 1.79]
Organization size (small)	0.55*** (0.13)	1.72 [1.34; 2.22]	0.02 (0.14)	1.02 [0.77; 1.35]	0.53*** (0.14)	1.70 [1.29; 2.24]
Organization size (mid-size)	0.15 (0.13)	1.16 [0.90; 1.50]	0.09 (0.14)	1.09 [0.83; 1.44]	0.06 (0.14)	1.07 [0.81; 1.41]
Autonomy	0.16*** (0.03)	1.18 [1.10; 1.25]	0.03 (0.04)	1.03 [0.96; 1.11]	0.13*** (0.03)	1.14 [1.07; 1.22]
Skills	0.14*** (0.04)	1.15 [1.07; 1.23]	0.14** (0.04)	1.15 [1.06; 1.25]	-0.00 (0.04)	1.00 [0.92; 1.08]
Security	0.28*** (0.04)	1.32 [1.23; 1.43]	0.11* (0.04)	1.11 [1.02; 1.21]	0.17*** (0.04)	1.19 [1.10; 1.30]
Stress	-0.25*** (0.04)	0.78 [0.73; 0.83]	-0.08* (0.04)	0.92 [0.86; 0.99]	-0.17*** (0.04)	0.85 [0.79; 0.91]
Importance of job	0.39*** (0.03)	1.47 [1.38; 1.58]	0.14*** (0.03)	1.15 [1.07; 1.22]	0.25*** (0.04)	1.29 [1.20; 1.39]
Job satisfaction	0.01** (0.00)	1.01 [1.00; 1.01]	0.02*** (0.00)	1.02 [1.01; 1.03]	-0.01*** (0.00)	0.99 [0.98; 0.99]

Notes. Reference group: the DRS class (left and middle part of the table), the semi-ERS class (right part of the table). R^2 =.13 (Nagelkerke). The model indicates a significant improvement in the fit comparing to the baseline model: $\chi^2(32) = 877.82$, $p < .001$.

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

Higher perceived job stress, on the other hand, makes an assignment to the ERS class less likely compared to the DRS class. In addition, assignment to the ERS class is more likely for employees in small organizations, and those who reported higher perceived autonomy, while it becomes less likely with higher education level and a high job position. The assignment to the semi-ERS class compared to the DRS class is more likely for female workers, and those who have a part-time job, and for workers with high job skills, high job security, less job stress, and higher importance of the job. The assignment to the ERS class, as opposed to the semi-ERS class, becomes less likely with a higher education level and increasing job stress. It becomes more likely for part-time employees and those who work in small organizations, and those who reported higher perceived autonomy, security, and importance of the job.

Apparently, individuals with a high probability of being assigned to the DRS class are male and full-time employees with higher perceived job stress and lower importance of the job. Individuals more likely assigned to the ERS class compared to the other two classes have a basic educational level, work part-time in small organizations and value their job as important. Both factors age and job satisfaction can hardly differentiate between the classes with different scale usage (the odds ratio are near to one).

2.6 Discussion

The popularity of measuring job satisfaction with an 11-point rating scale in national panel surveys contrasts with the lack of empirical research on the adequacy of such a long rating scale. The present study aimed to fill this gap and evaluate the adequacy of an 11-point rating scale. The application of a restricted mixed general partial credit model to JS data from the HILDA survey revealed severe drawbacks of an 11-point rating scale that we summarize in the next section. Thereafter, we will discuss the results of the multinomial logistic regression analysis that related the class-specific scale usage to covariates. Finally, we will discuss some implications and limitations of the study presented.

2.6.1 Drawbacks of an Eleven-Point Rating Scale

Low frequencies in specific categories

Explorative analysis showed the typical left-skewed distribution of the JS items, with very low frequencies in the first few categories. One can conclude that Australian employees are mostly satisfied with their jobs and hardly need to differentiate within the scale region pertaining to *dis*-satisfaction. The same can be observed for class-specific distributions of the JS items obtained from the model.

Patterns of inappropriate scale usage

Applying the rmGPCM to the JS items allowed us to identify three latent classes. About 40% of Australian employees were assigned to a class that avoids certain categories but differentiates reasonably among the remaining ones (up to six). In contrast, one-third of the sample was assigned to a class with extreme response style, dichotomizing the scale into the two extreme categories. The remaining class exhibited a pattern of differentiation at the extremes: The two lowest categories and the two highest categories were used. In general, we found that none of the latent classes differentiated between all of the 11 response categories proposed in the HILDA survey. Evidently, Australian employees evaluated their levels of job satisfaction using only two to six response categories.

Overall, these results are consistent with previous research that has assessed ISU for short rating scales (including four to six categories). However, our results add important aspects to existing knowledge, in particular with respect to the 11-point rating scale. This study detected a very high proportion of subjects with ISU. In fact, all latent classes subjectively reduced the number of response categories. Some form of the ERS was detected for 60% of the sample, while the proportion of ISU is commonly estimated to involve about a third of subjects. We detected two latent classes that used different forms of ERS. Previous studies, however, consistently reported one latent ERS class. Whereas the combination of avoided categories and the ERS has been previously observed (Eid & Rauber, 2000; Wu & Huang, 2010), the number of avoided categories was especially large in our study. The results revealed that researchers have to expect a larger number of unused categories and different forms of ERS the larger the number of response categories is.

Consequences for scale use

According to the results of our study, respondents use the scale in different ways. Therefore, comparing individuals using their total score may partly represent individual differences in response style and not differences in job satisfaction. It is important to note that more traditional psychometric methods are not able to detect these response style differences. For example, an exploratory factor analysis applied to the matrix of polychoric correlations of the JS five items indicated a one-factor solution (eigenvalues: 2.39, .83, .69, .65, .44). Furthermore, the psychometric quality of a scale may be overestimated when ISU effects are ignored. For example, after controlling for ISU by means of rmGPCM-3, the model-based reliability coefficient of the JS measure is lower ($\mathrm{Rel}_{\theta_v} = .59$) than the coefficient calculated on the basis of raw values of the JS items that contain ISU variance (Cronbach's $\alpha = .67$). According to the results presented researchers may obtain more valid results in studies aimed to explain individual differences in job satisfaction by taking response styles into account. This could be done by assigning individuals to response style classes first and then analyzing interindividual differences in job satisfaction by taking the estimated person parameters.

2.6.2 Explanation of Class Assignment

The pattern of relevant predictors distinguishing between the latent classes (see Table 2.4) encourages to think of a typical member of the respective classes. As mentioned above, members of the DRS class tend to be male, working full-time and to report high perceived job stress and low personal job importance. In contrast to the DRS class, a basic educational level and a low job position are related to the ERS class. Members of the ERS class evaluated their job conditions as positive (high job autonomy, suitable tasks with regard to qualifications, high job security, and low job-related stress). They tend to be female and working part-time, predominantly in small organizations. Similarly, members of the semi-ERS class are also more likely to be female and working part-time, but their educational levels do not differ from members of the DRS class.

2.6.3 Implications

For the population of Australian employees, a 6-point rating scale seems more adequate than the commonly used 11-point rating scale. We found that even in the differentiating class, only six of the proposed response categories mapped well onto the latent trait. Experimental studies that systematically vary the number of categories are needed to gather more evidence for an optimal number of response categories.

In general, mixed IRT models for polytomous data allow evaluating two aspects of ISU: avoided categories and response styles. Moreover, different types of ISU within a single sample can be identified. The results provide a researcher with relevant information concerning the appropriateness of the rating scale that was administered. In the majority of mixed IRT studies so far, researchers have been using the mPCM (and its restrictive variants) to detect ISU. In the present study, the rmGPCM exhibited a better fit to the data. Varying discrimination power of JS items seems a relevant item characteristic and should be taken into account. In contrast, for the JS data, a more general model such as the mGPCM-3 is suboptimal: Class-specific discrimination parameters are redundant, make the model too complex and reduce the estimation accuracy of parameters (as indicated by many extreme parameters and standard error estimates). In general, an application of mixed IRT models requires large sample size. If there are many parameters to estimate and the sample size is small, more restrictive models may be preferred over the rmGPCM to avoid estimation problems. Future simulation studies for mixed IRT models can help to clarify the optimal sample size required to obtain accurate parameter estimates under various data conditions (e.g., different scale lengths and rating scales).

2.6.4 Limitations

There were some minor estimation problems in the present application[8] that are most likely due to the low observed frequencies in the lower response categories.

While mixed IRT models for polytomous data applied for modeling ISU is classified as a typological approach, models of the dimensional framework (e.g., multidimensional IRT models, SEM) would be needed to quantify the intensity of ISU. The mixed IRT approach is beneficial for exploring ISU patterns in data, whereas multidimensional IRT models and SEM appear to be superior in eliminating effects of ad hoc-defined ISU from latent trait values (Wetzel, Böhnke, & Rose, 2015).

In this study, mostly contextual factors and socio-demographic variables were available to predict the ISU classes. Job conditions were found to be the best predictors. However, the full regression model was only able to explain a small portion of the variability in assignment to ISU classes (pseudo $R^2 =$ 13%). This re-emphasizes the need for research concerning causes of ISU. Future research should take into account the relative stability and trait-independence of ISU (e.g., ERS) and include cognitive ability (e.g., discrimination ability) and relevant dispositions (e.g., intolerance to ambiguity, decisiveness, impulsivity, social desirability). Because this is the first study that investigated ISU in job satisfaction data assessed using an 11-point rating scale, further studies are required to replicate these findings.

2.6.5 Conclusion

The present study is the first one that investigated the appropriateness of a long rating scale (an 11-point rating scale) for assessing aspects of job satisfaction by exploring ISU. Three scale usage patterns were extracted for Australian employees. Two features of class-specific ISU: (a) avoidance of several response categories in all latent classes and (b) preferred usage of the semi-ERS and ERS in two qualitatively different latent classes, provided empirical evidence that the 11-point rating scale contains redundant categories and evokes usage of simplification strategies. These findings show that a rating scale with 11 response categories is suboptimal for collecting high-quality data on job satisfaction and provide essential clues for more effective rating scales (2–6 categories; no middle point).

[8] The rmGPCM-3 contains one extreme threshold parameter and a few large standard errors for threshold parameters, mainly in the DRS class. A similar picture is obtained for the mPCM-3 (see Table 2.2).

2.7 References

Baumgartner, H., & Steenkamp, J.-B. E. M. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research, 38*, 143-156. doi: 10.1509/jmkr.38.2.143.18840

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52*, 345–370. doi: 10.1007/BF02294361

Carter, N. T., Dalal, D. K., Lake, C. J., Lin, B. C., & Zickar, M. J. (2011). Using mixed-model item response theory to analyze organizational survey responses: An illustration using the Job Descriptive Index. *Organizational Research Methods, 14*, 116-146. doi: 10.1177/1094428110363309

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*, 205-215. doi: 10.1177/014662169401800302

Cho, Y. (2013). The mixture distribution polytomous rasch model used to account for response styles on rating scales: A simulation study of parameter recovery and classification accuracy (Doctoral dissertation). Retrieved from http://drum.lib.umd.edu/bitstream/handle/1903/14511/Cho_umd_0117E_14472.pdf

Cox, E. P. (1980). The optimal number of response alternatives for a scale: A review. *Journal of Marketing Research, 17*, 407-422. doi: 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,* 201-225.

Eid, M., & Rauber, M. (2000). Detecting measurement invariance in organizational surveys. *European Journal of Psychological Assessment, 16*, 20-30. doi: 10.1027//1015-5759.16.1.20

Eid, M., & Zickar, M. (2007). Detecting response styles and faking in personality and organizational assessments by Mixed Rasch Models. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models* (pp. 255-270). New York: Springer Science + Business Media.

Enders, C. K. (2010). *Applied missing data analysis*. New York: Guilford Press.

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science, 8*, 206-213. doi: 10.1007/s11121-007-0070-9

Jin, K. Y., & Wang, W. C. (2014). Generalized IRT models for extreme response style. *Educational and Psychological Measurement, 74*, 116-138. doi: 10.1177/0013164413498876

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*, 376-407. doi: 10.1037/0033-2909.127.3.376

Kieruj, N. D., & Moors, G. (2013). Response style behavior question format dependent or personal style. *Quality & Quantity, 47*, 193-211. doi: 10.1007/s11135-011-9511-4

Kossek, E. E., & Ozeki, C. (1998). Work–family conflict, policies, and the job–life satisfaction relationship: A review and directions for organizational behavior–human resources research. *Journal of Applied Psychology, 83,* 139-149. doi: 10.1037/0021-9010.83.2.139

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,* 73-79.

Maij-de Meij, A. M., Kelderman, H., & van der Flier, H. (2008). Fitting a mixture item response theory model to personality questionnaire data: Characterizing latent classes and investigating possibilities for improving prediction. *Applied Psychological Measurement, 32*, 611-631. doi: 10.1177/0146621607312613

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174. doi: 10.1007/BF02296272

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*, 295-308. doi: 10.3758/BRM.41.2.295

Meiser, T., & Machunsky, M. (2008). The personal structure of personal need for structure: A mixture-distribution Rasch analysis. *European Journal of Psychological Assessment, 24*, 27-34. doi: 10.1027/1015-5759.24.1.27

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159-176. doi: 10.1002/j.2333-8504.1992.tb01436.x

Naemi, B. D., Beal, D. J., & Payne, S. C. (2009). Personality predictors of extreme response style. *Journal of Personality, 77*, 261-286. doi: 10.1111/j.1467-6494.2008.00545.x

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes.* (pp. 17-59). San Diego, CA US: Academic Press.

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*, 1-15. doi: 10.1016/S0001-6918(99)00050-5

Rost, J. (1997). Logistic mixture models. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 449-463). New York: Springer.

Spector, P. E. (1997). *Job satisfaction: Application, assessment, causes, and consequences*. Thousand Oaks, CA US: Sage Publications, Inc.

Summerfield, M., Freidin, S., Hahn, M., Li, N., Macalalad, N., Mundy, L., et al. (2015). *HILDA User Manual–Release 14*. Melbourne, Australia: Melbourne Institute of Applied Economic and Social Research, University of Melbourne.

Van Vaerenbergh, Y., & Thomas, T. D. (2013). Response styles in survey research: A literature review of antecedents, consequences, and remedies. Intern*ational Journal of Public Opinion Research, 25*, 195-217. doi: 10.1093/ijpor/eds021

Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political analysis, 18*, 450-469. doi: 10.1093/pan/mpq025

Vermunt, J. K., & Magidson, J. (2006). *Latent GOLD 4.0 and IRT modeling.* Statistical Innovations Inc, Belmont.

Vermunt, J. K., & Magidson, J. (2013). *Technical guide for Latent GOLD 5.0: Basic, Advanced, and Syntax.* Statistical Innovations Inc., Belmont.

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*, 108-124. doi: 10.1016/s0148-2963(01)00296-x

Von Davier, M., & Yamamoto, K. (2004). Partially observed mixtures of IRT models: An extension of the generalized partial-credit model. *Applied Psychological Measurement, 28*, 389-406. doi: 10.1177/0146621604268734

Weather, 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,* 1516-1524.

Weijters, B., Geuens, M., & Schillewaert, N. (2010). The individual consistency of acquiescence and extreme response style in self-report questionnaires. *Applied Psychological Measurement, 34*, 105-121. doi: 10.1177/0146621609338593

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*, 956-972. doi: 10.1177/0013164404268674

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*, 178-189. doi: 10.1016/j.jrp.2012.10.010

Wetzel, E., Böhnke, J. R., & Rose, N. (2015). A simulation study on methods of correcting for the effects of extreme response style. *Educational and Psychological Measurement, 76*, 304-324. doi: 10.1177/0013164415591848

Wetzel, E., & Carstensen, C. H. (2017). Multidimensional modeling of traits and response styles. *European Journal of Psychological Assessment., 33, 362-364.* doi: 10.1027/1015-5759/a000291

Wetzel, E., & Carstensen, C. H. (2014). Reversed thresholds in partial credit models A reason for collapsing categories? *Assessment, 21*, 765-774. doi: 10.1177/1073191114530775

Wu, P.-C., & Huang, T.-W. (2010). Person heterogeneity of the BDI-II-C and its effects on dimensionality and construct validity: Using mixture item response models. *Measurement and Evaluation in Counseling and Development, 43*, 155-167. doi: 10.1177/0748175610384808

2.8 Appendix to Chapter 2

2.8.1 Dimensional Structure of the Job Characteristics (JC) Measure

The HILDA survey (wave 2001) includes 12 JC items on aspects of job quality. But the HILDA user manual provides no information on evaluation of JC data. To examine the underlying structure of the JC measure, an explorative factor analysis (EFA) was conducted in Mplus software (available version 7, Muthen and Muthen, 2006) based on valid JC data of the analysis sample (*n* = 6302). We used the ML estimator. Factors were allowed to correlate (geomin rotation per default). The EFA detected four factors which eigenvalues are above 1 (eigenvalues: 2.85, 2.40, 1.61, 1.04, .90, .66, .58, .48, .43, .41, .36, .31). The standardized factor loadings of the four-factor solution are provided in Table 2.5. Most JC items could be assuredly assigned to only one factor (loadings above |.40| on one factor and loadings less than |.20| on others). The exceptions are two items: the item *c* 'I get paid fairly for the things I do in my job.' which has low loadings on all factors and the item *g* 'My job is complex and difficult.' which loads equally high on factor 1 and factor 4. These items were not considered by creating subscales. Thus, the four-factor structure reflexes the following aspects of job quality derived from the item content: (1) job-related stress, (2) job security, (3) degree of freedom in action and in decision-making at work, and (4) requirements of applying and extending one's professional skills.

Reference

Muthén, L. K., and Muthén, B. O. (2012). *Mplus: statistical analysis with latent variables; user's guide version 7*. Muthen & Muthen, Los Angeles, California.

Table 2.5. Standardized factor loadings in the exploratory factor analysis for Job Characteristics items.

Item	Factor 1	Factor 2	Factor 3	Factor 4
a. My job is more stressful than I had ever imagined.	**.82***	.03*	-.02*	.04*
b. I fear that the amount of stress in my job will make me physically ill.	**.77***	-.03*	-.02*	-.05*
c. I get paid fairly for the things I do in my job.	-.21*	.16*	.14*	-.02
d. I have a secure future in my job.	.06*	**.90***	.04*	.00
e. The company I work for will still be in business 5 years from now.	-.03*	**.55***	-.10*	.10*
f. I worry about the future of my job	.13*	**-.49***	.00	.08*
g. My job is complex and difficult.	.37*	-.00	.08*	.48*
h. My job often requires me to learn new skills.	.01	-.03*	-.08*	**.82***
i. I use many of my skills and abilities in my current job.	-.02	.07*	.18*	**.56***
j. I have a lot of freedom to decide how I do my own work.	-.06*	-.01*	**.82***	.04*
k. I have a lot of say about what happens on my job.	.03*	.06*	**.80***	.04*
l. I have a lot of freedom to decide when I do my work.	.01	-.05*	**.70***	-.10*

Notes. Factor loadings over |.40| are marked in bold. Items shown in gray are not assigned to any factors.

* *p* < .05.

2.8.2 Latent GOLD Script for Estimating the rmGPCM in Regression Submodule

```
options
   algorithm
      tolerance=1e-008 emtolerance=0.01 emiterations=8000 nriterations=600;
   startvalues
      seed=0 sets=100 tolerance=1e-005 iterations=200;
   bayes
      categorical=1 variances=1 latent=1 poisson=0;
   quadrature nodes=80;
   missing  excludeall;
   output
      parameters=first   // The first category parameter for all items is fixed to null.
      standarderrors=robust classification profile probmeans=posterior
      estimatedvalues=model iterationdetails identification
/* Estimated individual latent trait values and class assignment will be saved as a SAV file. */
      outfile 'estimated person parameters.sav' classification;
 variables
      caseid id;
      dependent response; // “response” consists of multiple responses of cases to all items.
      independent itemnr nominal; // “itemrt” is a nominal variable containing item IDs.
   latent
      theta continuous,
      class nominal 3; // “3” is the number of latent classes in the model.
 equations
      theta | class;   // Class-specific variance of the trait variable is freely estimated.
      class <- 1;      // Sum of class proportions is fixed to 1.
/* The model equation below is in the form of a logistic regression model with ’response’ as a dependent variable
and “itemnr” as an independent variable.
In this equation, the first term estimates differences in category parameters of two adjacent categories, which can
vary for all items in all classes. The second term in addition to the equation in the last line fixes the discrimination
parameter only of the first item to 1.*/
      response <- (~diff) 1 | itemnr class + (L) theta | itemnr;
      L[1]=1;

// For all latent classes, the mean of theta variable is fixed to null (default).
```

2.8.3 Missing Analysis

For checking missing data mechanisms, we compared groups with missing values and valid values on JC subscales using socio-demographic factors (age, gender, educational level, and total financial year income), personality (Big Five scales and personal control)[9], job-related variables (JC subscales, tenure in the current occupation, and job importance), and two latent variables (assignment to latent classes and estimated latent level of job satisfaction) obtained from the rmGPCM-3 application. We first built dichotomous variables for JC subscales (missing indicators). For continuous variables, we then conducted a series of independent-sample *t*-tests using each missing indicator as a group variable. Table 2.6 demonstrates *t*-test statistics and effect sizes (Cohen's standardized mean difference). For categorical variables, relative frequencies in the missing group and the χ^2-test statistics are reported. The non-significant test statistics indicate that missing data is completely at random (MCAR). In contrast, significant results and large effects indicate that other missing data mechanisms such as missing at random (MAR) or missing not at random (MNAR) are present (see Enders, 2010).

Table 2.6 includes only variables which produced significant test statistics.

References

Enders, C. K. (2010). *Applied Missing Data Analysis.* New York: Guilford Press.

Pearlin, L. I., and Schooler, C. (1978). The structure of coping. *Journal of Health and Social Behavior, 19*, 2-21.

Summerfield, M., Freidin, S., Hahn, M., Li, N., Macalalad, N., Mundy, L., et al. (2015). *HILDA User Manual–Release 14*. Melbourne, Australia: Melbourne Institute of Applied Economic and Social Research, University of Melbourne.

[9] The HILDA survey team first provides the data on the five dimensions of personality (e.g., extraversion, agreeableness, conscientiousness, emotional stability, and openness to experience) in the 5th wave (in 2005). The personal control scale (Pearlin and Schooler, 1978), which measures an extent of respondents' control over their life circumstances, is included in the 3rd wave (in 2003). (For details, see the HILDA User Manual – Release 14 by Summerfield et al., 2015). For this reason, we used these data only for handling missing values in predictor variables.

Table 2.6. Comparison of groups with missing and valid values for Job Charactersitics subscales.

	Comparison variable									
							Gender		Education	
Indicator	**Cons**	**Emot**	**PC**	**JS**	**Secu**	**MG**	**Male**	**Female**	**≤12 y.**	**> 12 y.**
Auto (*t*)	-4.57***	-1.95	-3.00^{1}**	-3,07^{1}**	-3.28**	%	9.3	7.5	9.5	7.7
(*d*)	-0.29	-0.13	-0.19	-2.65	-0.85	χ^2 (*df*)	6.94** (1)		6.71* (1)	
Secu (*t*)	-4.31***	-2.10*	-2.51^{1}*	-2.72^{1}**	-	%	10.0	8.7	10.8	8.4
(*d*)	-0.26	-0.13	-0.15	-2.14	-	χ^2 (*df*)	3.97* (1)		11.70** (1)	
Skills (*t*)	-4.53***	-1.85	-2.90^{1}**	-3.16^{1}**	-1.44	%	9.3	7.6	9.4	7.8
(*d*)	-0.29	-0.12	-0.18	-2.61	-0.36	χ^2 (*df*)	9.90** (1)		5.89* (1)	
Stress (*t*)	-4.37***	-1.86	-2.95^{1}**	-3.25**	-1.14	%	9.4	7.6	9.6	7.8
(*d*)	-0.28	-0.12	-0.18	-2.67	-0.25	χ^2 (*df*)	7.51** (1)		7.06** (1)	

Notes. Cons = Conscientiousness, Emot = Emotional stability, PC = Personal control, JS = estimated latent trait level of job satisfaction (scaled from 0 to 100), Auto = autonomy, Secu = security. *t* = T-Test Statistic. *d* = Standardized Mean Difference; *d* < 0 represents that the group with observed values had a higher mean, and *d* > 0 indicates that the persons with missing values had a higher mean. MG = missing group.

[1] Welch Statistic.

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

2.8.4 Goodness-of-Fit Statistics for the mPCM, rmGPCM, and mGPCM

Table 2.7. Goodness-of-fit statistics for the mPCM, rmGPCM, and mGPCM.

Model	N_{par}	$N_{\text{iter.}}$ in EM	$N_{\text{iter.}}$ in NR	LL	BIC	CAIC	AIC	BV	Extr. τ_{isg}	Extr. SE
mPCM										
1 class	51	1	2	-70131	140714	140765	140364	0	0	0
2 classes	103	2	2	-68555	138023	138126	137317	0	1	1
3 classes	155	2	2	-68173	137718	137873	136655	0	2	5
4 classes	207	2	3	-67943	137720	137927	136301	0	3	8
5 classes	259	2	6	-67796	137887	138146	136110	0	16	22
rmGPCM										
1 class	55	2	2	-69917	140321	140376	139944	0	0	0
2 classes	107	2	2	-68382	137712	137819	136979	0	0	0
3 classes	159	2	4	-68035	**137478**	**137637**	136388	0	1	7
4 classes	211	45	9	-67877	137622	137833	136175	0	4	25
5 classes	263	252	5	-67736	137802	138065	135998	0	10	23
mGPCM										
1 class	55	2	2	-69917	140321	140376	139944	0	0	0
2 classes	112	2	3	-68376	137745	137857	136976	1	28	1
3 classes	169	2	11	-68009	137515	137684	136356	2	98	7
4 classes	226	2	14	-67818	137639	137865	136089	3	266	21
5 classes	283	182	nc	-67692	137891	138174	135950	-	-	-

Notes. N_{par}: the number of model parameter. $N_{\text{iter.}}$ in EM: the number of iterations needed to reach convergence in EM algorithm. $N_{\text{iter.}}$ in NR: the number of iterations needed to reach convergence in Newton-Raphson algorithm. LL: Log-Likelihood. BIC: Bayesian information criterion. CAIC: Consistent Akaike's information criterion. AIC: Akaike's information criterion. Pearson *p*-Value: the bootstrapped *p*-value corresponding to the Pearson χ^2 goodness-of-fit statistic. CR *p*-Value: the bootstrapped *p*-value corresponding to the Cressie-Read χ^2 goodness-of-fit statistic. BV: boundary values. Extr. τ_{isg}: the number of threshold parameters larger than $|4|$. Extr. SE: the number of extreme standard errors of item parameters. (Extreme standard errors are defined as values five times larger than the most frequently occurring standard errors in the estimated model (here, larger than 1.5). nc: non-convergence. The lowest BIC and CAIC are marked in boldface.

3 Sample-Size Requirements for Applying Mixed Polytomous Item Response Models: Results of a Monte Carlo Simulation Study

Kutscher, T., Eid, M., & Crayen, C. (2018). *Sample-Size Requirements for Applying Mixed Polytomous Item Response Models: Results of a Monte Carlo Simulation Study*. Manuscript submitted for publication.

Abstract

Mixture models of item response theory (IRT) can be used to detect inappropriate category use. Data collected by panel surveys where attitudes and traits are typically assessed by short scales with many response categories are prone to response styles indicating inappropriate category use. However, the application of mixed IRT models to this data type can be challenging because of many threshold parameters within items. Up to now, there is very limited knowledge about the sample size required for an appropriate performance of estimation methods as well as goodness-of-fit criteria of mixed IRT models in this case. The present Monte Carlo simulation study examined these issues for two mixed IRT models (the restricted mixed generalized partial credit model [rmGPCM] and the mixed partial credit model [mPCM]). The population parameters of the simulation study were taken from a real application of survey research (5-item scale with an 11-point rating scale, three latent classes). For each model, data were generated based on varying sample sizes (from 500 to 5,000 observations with a 500-step). The effect of sample size on estimation problems and accuracy of parameter and standard error estimates were evaluated. Results show that the two mixed IRT models require at least 2,500 observations to provide accurate parameter and standard error estimates. The rmGPCM produces more estimation problems than the more parsimonious mPCM, mostly because of the sparse tables arising due to many response categories. These models exhibit similar trends of estimation accuracy across sample sizes. For model selection, the AIC3 and SABIC were the most reliable information criteria for a medium-sized sample (N = 1,500 and N = 2,500, respectively). The traditionally used BIC and CAIC only work well with large samples.

Keywords: mixture IRT models; rating scale; sample size; model selection; Monte Carlo simulation

Sample-Size Requirements for Applying Mixed Polytomous Items Response Models: Results of a Monte Carlo Simulation Study

Mixture models of item response theory (IRT) are a combination of IRT models and latent class analysis (see for an overview von Davier & Carstensen, 2007). They allow classifying individuals into homogeneous subpopulations that are a priori unknown (latent classes) and differ in the category characteristic curves linking the response probabilities with the latent trait variable (Rost, 1997). The mixture IRT approach for polytomous items is widely applied in empirical social research, mainly with the purpose of exploring population heterogeneity and its causes. For example, mixture polytomous IRT models are useful for detecting latent classes that qualitatively differ in a measured personality trait or attitude (e.g., Baghaei & Carstensen, 2013; Egberink, Meijer, & Veldkamp, 2010; Finch & Pierson 2011; Gnaldi, Bacci, & Bartolucci, 2016; Jensuttiwetchakul, Kanjanawasee, & Ngudgratoke, 2016) or those that are characterized by response styles (e.g., Austin, Deary, & Egan, 2006; Eid & Rauber, 2000; Eid & Zickar, 2006; Maij-de Meij, Kelderman, & van der Flier, 2008; Meiser & Machunsky, 2008; Wagner-Menghin, 2006; Wetzel, Carstensen, & Böhnke, 2013; Wu & Huang, 2010). Moreover, they can be applied to examine construct validity (e.g., Tietjens, Freund, Büsch, & Strauss, 2012; von Davier & Yamamoto, 2007), to detect differential item functioning (e.g., Cho, Suh, & Lee, 2016; Frick, Strobl, & Zeileis, 2015), and to check the quality of a rating scale in general (e.g., Kutscher, Crayen, & Eid, 2017; Smith, Ying, & Brown, 2011).

Compared to other statistical techniques that have been used to assess and control inappropriate category use (see for overview Van Vaerenbergh & Thomas, 2013), a distinguished advantage of the mixture IRT approach is that it can successfully represent different types of category use patterns (response styles) in one model. Mixed IRT models have been applied to detect response styles such as the preferences for extreme categories (ERS) or middle categories (MRS) (e.g., Maij-de Meij et al., 2008; Wetzel et al., 2013), faking or socially desirable responding (e.g., Mneimneh, Heeringa, Tourangeau, & Elliott, 2014; Ziegler & Kemper, 2013), and skipping superfluous response categories (e.g., Kutscher et al., 2017; Smith et al., 2011). Moreover, the application of a mixed IRT model does not require an a priori idea about the types of category use that may exist in the data, a single response style definition or an additional set of (heterogeneous) items in the questionnaire in order to measure response styles. Category use patterns are interpreted a posteriori based on the estimated class-specific item parameters. Due to its parsimony, the mixed partial credit model (mPCM; Rost, 1997) has been most often applied to explore category use in diverse research contexts (e.g., see Jasper, Nater, Hiller, Ehlert, Fischer, & Witthöft, 2013; Meiser & Machunsky, 2008; Wu & Huang, 2010). The assumption of equally discriminating items can be considered a disadvantage of the mPCM, because such data can hardly be observed in empirical reality, and if not met, such a restriction increases the probability of identifying a wrong number of latent classes

(Alexeev, Templin, & Cohen, 2011). Alternatively, the mixture extensions of multi-parameter IRT models (e.g., the generalized partial credit model [GPCM; Muraki, 1997] or the normal response model [NRM; Bock, 1972]) are more flexible and show a better fit to real-world data by including freely estimated discrimination parameters of items or categories (van der Linden & Hambleton, 1997). Only a few studies applied any of the latter group of models for exploring category use (see Egberink et al., 2010; Kutscher et al., 2017; Maij-de Meij et al., 2008). The hesitance to apply these models may partly be due to the lack of systematic research on the performance of complex mixture IRT models under various data situations (Embretson & Reise, 2013). For example, it is unclear whether an application of a complex mixed IRT model would require larger sample size or cause more estimation problems than a more parsimonious model.

To the best of our knowledge, only four simulation studies have examined the performance of (extended) mixture IRT models for polytomous items (excluding single-replication simulations), whose details are reported in Table 3.1. These are mixed one-, two- and three-parameter IRT models, some of which are extended by an additional class-specific parameter or random effect, allowing researchers to simultaneously unmix a sample into homogeneous latent classes and to control or quantify specific response-style effects. In general, the simulation conditions of these studies included varying sample sizes (200 up to 6,000 observations), scale lengths (4 up to 50 items), response formats (with 4 up to 6 ordered response categories), and features of latent classes (e.g., the number of latent classes, class sizes). These simulation studies focused on applying mixed IRT models for the purpose of individual diagnostic and obtaining accurately estimated individuals' trait values when latent heterogeneity of a target population and effects of category use are taken into consideration. It is well known that IRT models require sufficiently long scales to precisely estimate individuals' trait values (DeMars, 2003; He & Wheadon, 2013; Kieftenbeld & Natesan, 2012; Meyer & Hailey, 2012; Reise & Yu, 1990). In these simulation studies the items showed only a few number of response categories to prevent estimation problems (DeMars, 2003; Choi, Cook, & Dodd, 1997; De Ayala & Sava-Bolesta, 1999; De la Torre, Stark, & Chernyshenko, 2006; French & Dodd, 1999; He & Wheadon, 2013; Lange, 2008). Hence, all these simulation studies are characterized by (relatively) long scales (10 to 50 items) and few response categories (4- to 6-point rating scales).

However, these simulation studies have hardly included the data situation that is often observed in national panel surveys and large-assessment surveys where the measurement of attitudes and traits are based on short scales (e.g., the 5-item measure of job satisfaction in the Household, Income and Labour Dynamics in Australia survey [HILDA; Summerfield, Freidin, Hahn, Li, Macalalad, Mundy, … & Wooden, 2015]; the 5-item measure of satisfaction with working condition in Swiss Household Panel study [SHP; Voorpostel, Tillmann, Lebert, Kuhn, Lipps, Ryser, ... & Wernli, 2014]). Clearly, in context of panel studies, it is impractical to use long-scale measures, primarily to keep the time required to

Table 3.1. Overview of the simulation studies on the performance of mixture polytomous IRT models.

Model description	Design	Main finding and acceptable data condition
Huang (2016) Two mixed GPCMs with a random-effect RS-variable (the so-called mixture ERS-GPCM[1] and the mixture ERS-GPCM-CD[2])	Fixed factors: - Latent mixture: 3 classes (ORS class [50%], ERS class [25%], and MRS class [25%]) Varied factors: - Sample size: 200; 500; 1,000; and 2,000 cases - Scale length: 10, 15, 20, and 40 items - Rating scale: 4 and 6 categories Bayesian estimation method	Optimal performance: - *N* = 1,000 cases and 20 items. Further relevant results: - Accuracy of parameter estimates and classification rates are positively associated with longer scales, larger sample sizes, and more response options; - More accurate parameter estimates in the larger class than in small classes; - High non-convergence rate in the case of short scales and small sample sizes.
Jin and Wang (2014) The mixed 3P-GPCM with class-specific decrement parameter[3]	Fixed factors: - Sample size: 2,000 respondents - Scale length: 20 items - Rating scale: 4 categories - Unequal class sizes: 60% and 40% Varied factors: - Latent mixture: 1 and 2 classes - Decrement parameter: 0, .1, and .2 Bayesian estimation method	- Optimal parameter recovery under all simulation conditions (*RMSE* < .11).
Wetzel, Böhnke, and Rose (2016) The mixed PCM (mPCM; Rost, 1997)	Fixed factor: - Rating scale: 4 categories Varied factors: - Latent mixture: 1 class and 2 classes (ERS class [50%] and NERS class [50%]) - Sample size: 200, 500, and 2,000 cases - Scale length: 5, 10, 25, and 50 items Marginal maximum likelihood (MML) estimation method	- For the one-class-PCM, high recovery accuracy of person parameters with 10 or more items across all sample sizes and scale lengths. - For the two-class-mPCM, the mean probabilities of class membership are high for all scale lengths; moderate accuracy of person parameters for the short scale (5 items) across all sample sizes; high accuracy of person parameters for scales with more items.

Model description	Design	Main finding and acceptable data condition
Cho (2014) The mixed PCM (mPCM; Rost 1997) with latent classes representing RSs (ORS, ERS, MRS, or ARS)	Fixed factor: - Rating scale: 5 categories Varied factors: - Sample size: 1,200; 3,000; and 6,000 cases - Scale length: 4, 10, and 20 items - Latent mixture: 2, 3, and 4 classes - Class sizes: equal and unequal (the ORS class as a large one and each other RS class consists of 10% of the sample) Marginal maximum likelihood (MML) estimation method	Optimal performance: - For the four-class mPCM (equal classes), *N* = 3,000 cases and 10 items. Further relevant results: - The mPCM with fewer classes required less than *N*=3,000 respondents. - For unequal classes, more cases are needed to achieve the same accuracy compared to equal classes. - Class-specific parameters of small classes showed less accurate estimates. - The test length was the main factor affecting the accuracy of ability parameter recovery. - The test length was the most important factor for classification accuracy, regardless of the sample size. - A higher misclassification rate for small classes and classes with similar class-specific item parameters (e.g., ERS class and ARS class). Estimation problems (non-convergence and boundary values): - For the four-class mPCM (unequal classes), with small sample size and short test length, mostly due to insufficient expected category-frequencies (near zero), especially for in small classes.

Notes. GPCM: Generalized partial credit model. RS: response style. ORS: ordinary response style. ERS: extreme response style. MRS: middle response style. NERS: non-extreme response style. ARS: acquiescence response style, *RMSE*: the root mean squared error. PCM: Partial credit model. mPCM: mixed partial credit model.

[1] The so-called mixture ERS-GPCM allows to detect latent classes with different response patterns and additionally quantify an individual tendency for ERS. For this purpose, it includes an additional random-effect factor that represents interindividual differences in category widths.

[2] The so-called mixture ERS-GPCM-CD is an extension of the mixture ERS-GPCM and includes an additional item-specific constrained discrimination (CD) parameter. It makes possible to identify items that strongly evoke ERS.

[3] The 3P-GPCM with class-specific decrement parameter is the most complex extension of the mixed GPCM. It includes a decrement parameter which allows to quality a possible decline in respondents' response behavior (because of a time limit, low motivation or insufficient ability).

respond to the questionnaire within a reasonable limit to prevent any reduction in participants' motivation and to collect data of high quality. Moreover, short scales are usually compensated by a rating scale consisting of many response categories (e.g., an 11-point rating scale) with the purpose to measure fine gradations of individuals' trait levels on a trait or an attitude of interest (see Krosnick & Presser, 2010; Willits, Theodori, & Luloff, 2016). However, empirical research has shown that scales with many response categories are affected by reduced psychometric data quality due to increased error variance as a consequence of response styles evoked by many categories (Chang, 1994; Weng, 2004). It is precisely this data situation which makes the use of mixed IRT models particularly reasonable, enabling a researcher to explore category use patterns existing in the data and to adjust estimates of individuals' latent trait values. Thus, the present simulation study focuses on examining under which conditions (namely, sample size) mixed IRT models for polytomous items would perform appropriately when they are applied to data assessed with a short scale (5 items) and many response categories (11 response categories).

3.1 Determining the Number of Latent Classes Using Information Criteria

One critical issue in applying mixed IRT models is the determination of the number of latent classes. This is typically done by applying information criteria (e.g., the Akaike's information criterion [AIC; Akaike, 1974], the Bayesian information criterion [BIC; Schwarz, 1978] or consistent AIC [CAIC; Bozdogan, 1987]). Because information criteria are differently affected by the model complexity (the number of model parameters) and sample size, they usually provide inconsistent suggestions concerning the best-fitting class solution (Cho, 2014; Choi, Paek, & Cho, 2017; Li, Cohen, Kim, & Cho, 2009; Yu & Park, 2014). Therefore, the conditions under which these information criteria perform well have to be explored. Several simulation studies have been conducted to give an answer to this question.

In his extensive simulation study, Cho (2014) examined the effectiveness of traditional information criteria such as the AIC, BIC, and CAIC for determining the true number of latent classes of the mPCM under different simulation conditions. He found that the BIC generally performed well in the most conditions, followed by the CAIC showing a slightly lower overall accuracy rate. In contrast, the asymptotically inconsistent AIC often overestimated the true number of latent classes, especially with larger sample sizes. Consistent findings have also been reported for mixture dichotomous IRT models (Cho, Cohen, & Kim, 2013; Li et al., 2009; Preinerstorfer & Formann, 2012). However, Cho (2014) concluded that the BIC and CAIC are not the best. For example, both information criteria tend to underestimate the true number of latent classes in the case of insufficient sample size (e.g., fewer than 1,000 observations) and when complex mixture models are applied (Bozdogan, 1987; Cho, 2014; Choi et al., 2017; Dias, 2006; Nylund, Asparouhov, & Muthén, 2007; Yang & Yang, 2007; Yu & Park, 2014). In these conditions, the AIC performs better.

In other studies, the AIC whose penalty term includes triple the number of model parameters (AIC3; Bozdogan, 1994) and the sample-size adjusted BIC (SABIC; Sclove, 1987) have been proven to overperform the BIC, CAIC, and AIC, especially for relatively small sample sizes (Andrews & Currim, 2003; Choi et al., 2017; Dias, 2006; Fonseca, 2010; Nylund et al., 2007; Yang & Yang, 2007; Yu & Park, 2014). The AIC3 can detect the true latent mixture structure with a high accuracy rate (above 90%) almost regardless of the sample size if it consists at least of 500 respondents (Fonseca, 2010; Yang & Yang, 2007). In contrast to the BIC, the SABIC, which less penalizes the model complexity, showed a lower underfitting rate under reasonably small sample sizes (Choi et al., 2017). Both the AIC3 and SABIC were proper in detecting complex latent mixtures with more than two classes (Yang & Yang, 2007; Yu & Park, 2014). Although these simulation studies provide important insight into the appropriateness of different information criteria, it is unknown whether they behave appropriately in the context that is typical for survey research (short scales, long rating scales).

3.2 Objectives of the Study

The objective of this study is to examine the required sample size for two mixed polytomous IRT models that are primarily used for exploring category use by means of Monte Carlo simulations. The restricted mixed GPCM (rmGPCM; with varying discrimination parameters of items only for the total population but not across latent classes) and the mPCM (with equal discrimination parameters of items) are compared in their performance under small to large sample sizes. Both models have been well established in research on category use (e.g., Austin et al., 2006; Eid & Rauber, 2000; Kutscher et al., 2017; Meiser & Machunsky, 2008; Wetzel et al., 2013). The study primarily focuses on realistic data situations in the field of national surveys, where psychological constructs are assessed using short scales with many response categories. To prevent the main limitation of previous simulation studies, we use empirically-based model parameters reflecting the latent mixture of three subpopulations with different category use patterns. Because Monte Carlo simulation studies on mixture IRT models are very time consuming, only the sample size is manipulated, whereas further factors (such as scale length, the number of response categories, type of latent mixture, and proportions of latent classes) are fixed. In particular, differences in estimation accuracy for the rmGPCM and mPCM under varying sample sizes are investigated. Furthermore, we compare different information criteria in their performance for correctly identifying the true class solution of both models. This study should provide an insight into requirements and obstacles when exploring category use by means of the mixed IRT models in the presence of a challenging data situation (5 items with 11 response categories). To the best of our knowledge, this is the first study investigating the mixed one- and two-parameter IRT model for polytomous data in the context of a short scale and a large number of response categories and, therefore, will add a valuable contribution to the literature on the mixture IRT approach.

3.3 Method

3.3.1 Data-Generating Models

In the current simulation study, we use the rmGPCM and the mPCM (Rost, 1997) as data generating models. As a parsimonious variant of the mixed GPCM (GPCM; Muraki, 1997; mGPCM; von Davier & Yamamoto, 2004), the rmGPCM defines for each latent class the conditional probability of endorsing response category *x* of item *i* as a function of the latent trait variable by two types of item parameters: (i) class-specific threshold parameters that define the location of transition between two adjacent categories of item *i* (*x* - 1 and *x*) on the latent continuum and (ii) a class-fixed discrimination parameter of item *i* (as a multiplicative parameter) that indicates how well an item differentiates between individuals with different values on the trait measured. That means that the location and the order of thresholds can differ between latent classes. The discrimination parameters are freely estimated for items and are fixed across latent classes. The rmGPCM is defined by the following equation:

$$P_{vix}(\theta) = \sum_{g=1}^{G} \pi_g \frac{\exp\,[\sum_{s=0}^{x} \delta_i(\theta_{vg} - \tau_{isg})]}{\sum_{c=0}^{m} \exp\,[\sum_{s=0}^{c} \delta_i(\theta_{vg} - \tau_{isg})]} \tag{3.1}$$

with $x \in \{0,\ldots, m\}$, $s \in \{0,\ldots, c\}$, $\delta_i > 0$; $\sum_{g=1}^{G} \pi_g = 1$, $\mathrm{E}(\theta_{vg}) = 0$ for all *g*, $\tau_{i0g} = 0$ for all *i* in all *g*, $\delta_1 = 1$ (as identification constraints).

In Equation 1, the proportion of individuals in each latent class (class sizes) π_g ($0 < \pi_g < 1$), the class-specific threshold parameters for item *i* (τ_{isg}), the item-specific discrimination parameters (δ_i), and the class-specific values on the latent trait which are measured for person *v* (θ_{vg}) are all model parameters to be estimated. $P_{vix}(\theta)$ denotes the probability of individual *v* endorsing category *x* of item *i*. The number of a priori unknown subpopulations (*G*) can be determined by comparing goodness-of-fit statistics of models differing in the number of latent classes (Rost, 1997). In addition, the class membership *g* (*g* = 1,…, *G*) of each individual can be determined by his or her maximal class assignment probability. Mathematically, the mPCM (Rost, 1997) is a special variant of the rmGPCM. This model assumes that the discrimination parameters do not differ between items and classes and are usually fixed to one. In both models, the threshold parameter values have the same meaning.

In the present simulation study, data generating and data analysis were implemented in the Latent GOLD 4.5 package (Vermunt & Magidson, 2008). It should be noted that in this software the parametrization of mixed IRT models is based on the generalized linear model (GLM), and, therefore, model parameters are partially generated in a different metric as commonly used in the IRT approach (e.g., difference of adjacent category parameters instead of threshold parameters). For example, the model equation for the rmGPCM has the following form of a logistic regression model:

$$log \frac{P(Y_i = m|F_v,g)}{P(Y_i = m-1|F_v,g)} = (\beta^i_{0mg} - \beta^i_{0m-1g}) + \lambda^i F_v, \tag{3.2}$$

where Y_i is an observed response for item *i* and, F_v is a person's latent trait value (representing the weighted average of one's class-specific ability parameters), $(\beta^i_{0mg} - \beta^i_{0m-1g})$ denotes a parameter for the difference of category difficulty parameters of two adjacent categories *m* and *m* - 1 for item *i* in the class *g* (the so-called delta beta parameter, $\Delta\beta^i_{0sg}$), and λ^i is an item discrimination parameter. The results are reported with respect to the Latent GOLD parameterization.

3.3.2 Simulation Design

The present simulation study examined what sample size is required to avoid estimation problems and to obtain accurately estimated model parameters, standard errors and correct model fit coefficients for the specific data condition characterized by a short scale and a large number of response categories.

Two factors were manipulated in the simulation study: (i) model type (the rmGPCM, the mPCM) and (ii) sample size (starting from 500 observations up to 5,000 observations with a step of 500 observations). Sample sizes were chosen to represent realistic data conditions. These two manipulated factors were crossed resulting in 20 simulation conditions.

The data-generating models used in the current study were the three-class rmGPCM and the three-class mPCM (described in their general form in the previous section). The generating parameter values of both models (taken as population parameters) were drawn from an application to empirical survey data reported by Kutscher and colleagues (2017) and are shown in Table 3.2. In this empirical application, five items measuring job satisfaction on an 11-point rating scale from the first wave of the Household, Income and Labour Dynamics in Australia survey (Summerfield et al., 2015) were analyzed. Fitting the data with both models, three latent classes with different category use were detected based on a subsample of 7,036 employees and employers. In this application, the three-class rmGPCM showed the best fit. The three classes can be characterized as follows: The first class shows an ERS with a large number of avoided categories (indicated by many unordered thresholds); the second class is characterized by a roughly ordinary response style (ORS) and a few avoided response categories (indicated by approximately equal widths between adjacent threshold parameters and a few unordered thresholds); members of the third class prefer the two lowest and two highest response categories (semi-ERS) with

Table 3.2. Generating parameters of the rmGPCM-3 (upper lines) and mPCM-3 (bottom lines).

	λ^i	$\Delta\beta^i_{01g}$	$\Delta\beta^i_{02g}$	$\Delta\beta^i_{03g}$	$\Delta\beta^i_{04g}$	$\Delta\beta^i_{05g}$	$\Delta\beta^i_{06g}$	$\Delta\beta^i_{07g}$	$\Delta\beta^i_{08g}$	$\Delta\beta^i_{09g}$	$\Delta\beta^i_{010g}$	λ_g	π_g
						Class 1							
Item 1	1[1]	-.98	.77	.52	.08	1.22	-.61	.78	.24	-1.24	1.89		
	1	-.87	.85	.58	.16	1.25	-.58	.80	.24	-1.26	1.86		
Item 2	.71	-.79	.71	-.07	.03	1.26	-1.04	.56	.96	-.69	2.49		
	1	-.59	.94	.09	.20	1.38	-.94	.64	1.02	-.70	2.43		
Item 3	1.27	-.28	1.07	.56	.07	1.76	-.32	.81	.62	-.94	2.05	.21	0[1]
	1	-.20	.98	.59	.09	1.76	-.29	.80	.64	-.97	2.08	.28	0[1]
Item 4	2.57	-.31	1.98	.66	.39	1.97	-.55	.88	.75	-1.43	2.27		
	1	-.74	1.39	.23	.00	1.72	-.81	.76	.72	-1.39	2.45		
Item 5	1.76	-.52	1.00	.32	.28	1.63	-1.12	1.21	.67	-.95	2.43		
	1	-.67	.73	.18	.18	1.53	-1.27	1.21	.67	-.95	2.54		
						Class 2							
Item 1	1[1]	2.27	1.52	.86	.40	.64	.25	.46	.10	-1.92	-1.16		
	1	2.01	1.39	.94	.42	.74	.20	.48	.06	-1.66	-1.32		
Item 2	.71	1.72	1.21	.83	-.03	.82	.04	.59	.70	-.67	-.21		
	1	1.96	1.34	1.05	.11	.98	.06	.70	.67	-.55	-.25		
Item 3	1.27	-3.10	4.00[2]	1.80	.58	1.05	.38	.68	.45	-1.38	-1.20	.24	.20
	1	.63	2.78	1.58	.54	1.02	.34	.68	.42	-1.20	-.1.01	.30	.30
Item 4	2.57	4.00[2]	3.28	1.70	1.07	1.05	.28	.59	.06	-2.21	-1.73		
	1	4.00[2]	2.02	1.10	.55	.76	.03	.51	.12	-2.12	-1.86		
Item 5	1.76	1.91	1.86	.98	.38	.84	.30	.59	.29	-.99	-.93		
	1	1.49	1.43	.73	.19	.68	.15	.55	.27	-.96	-.98		
						Class 3							
Item 1	1[1]	1.61	.60	.35	-.22	.83	.01	.85	.03	.11	-1.93		
	1	1.73	.78	.40	-.15	.80	.17	.84	.04	.01	-1.89		
Item 2	.71	.27	.20	-.43	.48	1.01	-.94	1.58	.23	.96	-.66		
	1	.47	.44	-.35	.71	1.07	-.69	1.46	.27	.86	-.74		
Item 3	1.27	2.27	.60	.34	-.01	1.06	.33	.87	.20	.96	-1.46	.21	-.18
	1	2.27	.59	.29	-.13	1.26	.37	.91	.29	.93	-1.53	.28	-.25
Item 4	2.57	4.00[2]	.65	.79	.22	1.51	.21	.63	.12	.77	-2.03		
	1	4.17	.13	.16	-.34	1.56	.18	.65	.33	.95	-1.68		
Item 5	1.76	1.15	.68	.26	.32	1.04	-.32	.63	.54	.78	-.70		
	1	.90	.51	-.13	.29	1.27	-.37	.63	.68	.94	-.56		

Notes. λ^i: discrimination parameter. $\Delta\beta^i_{0sg} = (\beta^i_{0mg} - \beta^i_{0m-1g})$: delta beta parameter. λ_g: estimate of the variance of the class-specific latent trait variable. π_g: latent class-size parameter. All parameters are given in Latent GOLD metric (see Vermunt & Magidson, 2006).

[1] Default setting. [2] Extreme parameters were substituted by |4|.

many avoided categories between.

Because the generating model parameters are based on the empirical study (to guarantee the ecological validity of the simulation study), the following four factors were fixed under different sample-size conditions in the present simulation study: (i) the length of the scale (5 items); (ii) the number of response categories (an 11-point rating scale); (iii) the number of latent classes (3 classes with different category use patterns described above), and (iv) the sizes of latent classes (for the three-class rmGPCM: 0.33, 0.40 and 0.27; for the three-class mPCM: 0.32, 0.43 and 0.25). In this simulation study, 500 replications were generated per condition. Within a replication, we estimated also the one- to four-class solutions of a corresponding model.

In the present study, the marginal maximum likelihood (MML) estimator implemented in Latent GOLD was used for estimation of both models. For effective MML estimation, the stable EM algorithm (Bock & Aitkin, 1981) is used in the initial stage of the estimation process and it switches to the speedy Newton-Raphson (NR) method in the final stage (Vermunt & Magidson, 2013). Each estimation algorithm stops when its maximum number of iterations or the convergence criterion (equals to .01 and to 10^{-8} per default for the EM algorithm and the NR algorithm, respectively) is reached. In order to prevent estimation problems (such as non-convergence or local maximum), the following estimation options were chosen, for all class solutions in all sample-size conditions: (i) the number of iterations for the EM algorithm and for the NR method were fixed to 10,000 and to 600, respectively; (ii) the number of multiple sets of starting values was set to 100 and the number of EM iterations performed within each start set was set to 200; (iii) following Muraki's suggestion (1997), the number of quadrature points was set to 80. Further options were retained to be Latent GOLD default values.

3.4 Analyses

3.4.1 Monitoring Convergence and Estimation Problem

To evaluate the estimation performance of the rmGPCM and mPCM, convergence checks were conducted for each analysis of each replication by considering the convergence rate of the EM algorithm and the NR estimation method and the occurrence of boundary values. Latent GOLD indicates these estimation problems with warning messages. Consequently, replications with warning messages were inspected. A high rate of boundary values within a class solution (e.g., over 10%) is indicative of an improper solution. Replications with an improper three-class solution of rmGPCM and mPCM were eliminated from sequential analyses (for details on this issue, see Results section).

3.4.2 Detection of Label Switching

Evaluating the estimation accuracy of the three-class rmGPCM (rmGPCM-3) and the three-class mPCM (mPCM-3) across replications requires the match in the order of latent classes between the data-generating model and replications (exclusion of label switching). A useful approach to detect switched classes within a replication is based on comparing class-specific item parameters used for data generating with the estimates from each replication (see Cho, 2014; Cho et al., 2013; Li et al., 2009).

In the present simulation study, label switching should actually be prevented by using data-generating parameters as starting values for estimating three-class rmGPCM and the three-class mPCM in corresponding replications (Vermunt & Magidson, 2016). To ensure that it had worked well, we checked the occurrence of switched classes by means of the multinomial logistic regression analysis within each condition. This method was based on fifty delta beta parameter estimates from each of replications and predicted their assignment to a certain latent class. In all conditions, a perfect correspondence between observed and predicted class assignments of delta beta parameters was found (complete separation). Hence, as expected, no label switching occurred.

3.4.3 Measures of Estimation Accuracy

The estimation accuracy was evaluated using the following robust accuracy indices: Root median squared error (*RMdSE*), standard error bias (*bias*$_{\mathrm{se}}$), median width of the confidence interval ($Md_{\mathrm{width}_{CI}}$), Spearman's rank correlation coefficient (r_{s}), and 95% coverage. We primarily used the median-based measures that are robust to extreme estimates that may occur as a consequence of sparse data.

The *RMdSE* is a robust measure of absolute accuracy of parameter estimation computed by

$$RMdSE = \sqrt{Md_{(\hat{\mathrm{p}}-\mathrm{p})^2}} \tag{3.3}$$

where $\hat{p}$ denotes the parameter estimate of the t^{th} replication and p represents the generating parameter value. Thus, this index is based on squared differences between estimated and the true parameters each of those is calculated for a replication; the squared root of the median is then used to aggregate these differences across replications. The fewer parameter estimates across replications deviate from the true parameter value, the smaller is the *RMdSE* observed.

The standard error bias (*bias*$_{\mathrm{se}}$) demonstrates how well the standard error of a parameter is reproduced by the standard deviation of the empirical distribution of its estimates across replications. Thus, standard error bias was calculated as the median of absolute differences between standard error ($\widehat{se}$) estimate of a parameter in t^{th} replication and empirical standard deviation ($SD_{\hat{\mathrm{p}}}$) of parameter estimates across all replications:

$$bias_{\mathrm{se}} = Md_{|\widehat{\mathrm{se}} - SD_{\widehat{\mathrm{p}}}|}. \tag{3.4}$$

If standard error estimates of a parameter are close to the empirical standard deviation of the parameter distribution, the $bias_{\mathrm{se}}$ should be close to zero.

The median width of the confidence interval ($Md_{\mathrm{width}_{CI}}$) is a robust measure of the estimation accuracy of standard errors. Small standard errors affect narrow confidence intervals and thus indicate accurate parameter estimation. For a parameter in the t^{th} replication, the width of a 95% confidence interval was calculated using the estimated standard error and the 97.5th quantile of the standard normal distribution; then, the median was used to aggregate these statistics across replications as follows:

$$Md_{\mathrm{width}_{CI}} = Md_{(2*\mathrm{z}_{(.975)}*\,\widehat{\mathrm{se}})} \tag{3.5}$$

To obtain only one statistic for the *RMdSE*, $bias_{se}$, and $Md_{width_{CI}}$ across delta beta parameters in latent classes, all calculated accuracy indices were aggregated based on the median. Similarly, the average coverage was calculated, separately for latent classes. Before calculating the estimation accuracy indices, extreme parameter estimates (>|10|), extreme standard error estimates (> 50), and boundary values of standard errors, including their corresponding standard errors and parameters, were eliminated (for details, see Result section).

In addition, for class-specific delta beta parameters of each item, Spearman's rank correlation (r_{S}) was calculated between the population parameters and their estimates within a replication. It provides how accurately these estimates represent a class-specific response pattern which is inherent in the data-generating parameters. Correlation coefficients were then averaged across replications and items. A high average correlation coefficient (at least .90) demonstrates the highly concordant order of estimated and generating delta beta parameters in latent classes.

Finally, the 95% coverage was calculated that reflects the proportion of replications for which 95% confidence interval covers the generating parameter value. Coverage of at least .90 is optimal. All analyses were performed using R 3.3.0 (R Development Core Team, 2016).

3.4.4 Detection of the True Class Solution

The current study evaluated how effective five information criteria implemented in Latent GOLD are for identifying a true class solution for applications of the rmGPCM and mPCM. The information criteria considered are defined as follows:

$$\text{AIC} = -2\text{LL} + 2*\text{N}_{\text{par}} \tag{3.6}$$

$$\text{BIC} = -2\text{LL} + \log(\text{N})*\text{N}_{\text{par}} \tag{3.7}$$

$$\text{CAIC} = -2\text{LL} + [\log(\text{N}) + 1]*\text{N}_{\text{par}} \tag{3.8}$$

$$\text{AIC3} = -2\text{LL} + 3*\text{N}_{\text{par}} \tag{3.9}$$

$$\text{SABIC} = -2\text{LL} + [\log(\frac{N+2}{24})]*\text{N}_{\text{par}} \tag{3.10}$$

where -2LL is -2 times the log-likelihood of the class solution, N_{par} is the number of parameters to be estimated, and N is the sample size.

The class solution with the smallest value of an information criterion is indicated as the best-fitting model. In the present study, for each information criterion coefficient, the proportion of replications in which a specific class solution of the rmGPCM or mPCM was identified as the best-fitting model was calculated and compared under different sample-size conditions. We considered an information criterion as appropriate when it could correctly identify the true three-class solution for at least 95% of replications generated by the corresponding mixed IRT model.

3.5 Results

3.5.1 Convergence and Estimation Problems

Table 3.3 gives an overview of convergence and estimation problems for the rmGPCM-3 and mPCM-3. For the rmGPCM-3, the EM algorithm converged in all replications. Contrarily, the convergence rate of the rapid NR algorithm reached only 69% of replications across all conditions and this is considerably reduced with increasing sample size (from 84% to 56% with *N* = 500 and *N* = 5,000, respectively). Coincidently, boundary estimates occurred in almost all non-convergent replications. A detailed analysis revealed that the boundary values problem mostly referred to the standard error estimates of the same delta beta parameters $\Delta\beta_{02,g=2}^{i=3}$ (in 83% of the non-convergent replications). Note that in the empirical application of the rmGPCM-3 to the HILDA data (population model), this parameter was estimated to be extreme (see Table 3.2) because the expected frequencies of two lower categories of item 3 in the second class were null (see Table 3.13 in the appendix to Chapter 3). Obviously, a sparse table seems to be a challenge for the NR method. By increasing the sample size, the high rate of boundary estimates of the standard error concerned (and consequently that of non-convergent replications) may be explained by the fact that these adjacent response categories still did not provide sufficient data points required for accurate estimation of this delta beta parameter by the NR algorithm in the certain *N* condition. Furthermore, seven replications across all conditions were identified as improper (as reported in parentheses in the column "BV$_{SE}$" of Table 3.3) and were completely excluded from the subsequent analyses.

The mPCM-3 showed more satisfactory results (see the bottom part of Table 3.3). The EM algorithm also converged in all sample-size conditions. The non-convergence rate of the NR algorithm was maximal 4% and concerned only sample-size conditions with fewer than 3,000 observations. This was mostly combined with the occurrence of the standard error of (extreme) delta beta parameters indicating a boundary value. No improper solutions were found. Regarding other class solutions of the two models, the same non-convergence and estimation problems in a greater extent were found for the four-class solutions (see Tables 3.7 – 3.8 in the appendix to Chapter 3). To conclude, the NR algorithm can fail to achieve a convergent solution in a case of a high model complexity (e.g., mixed multi-parameter IRT model, many latent classes) and in the presence of sparse data.

3.5.2 Accuracy of Estimates

At first, we examined the presence of extreme values of parameter and standard error estimates in the three-class solutions. For the rmGPCM-3, a total of 3% of the parameters and a few standard errors (0.03%, excl. boundary values) were estimated to be extreme in all replications across all sample-size conditions. Mostly, it referred to four delta beta parameters ($\Delta\beta_{01,g=3}^{i=4}$, $\Delta\beta_{02,g=2}^{i=4}$, $\Delta\beta_{01,g=2}^{i=3}$, $\Delta\beta_{01,g=2}^{i=4}$) whose values in the population model are also relatively high (see Table 3.2). The standard errors of the first and second delta beta parameters often obtained extreme values, primarily because of the sparse table problem occurring in the presence of lower response categories (see Table 3.13 in the appendix to Chapter 3). For the same reasons, the mPCM-3 produced a few extreme values of parameter estimates (0.01%) and standard errors (0.4%) across all replications. All extreme estimates, boundary values, and their corresponding parameters and standard errors were excluded from the following analysis. Below, we will report relevant results separately for accuracy indices. (Exact values on accuracy indices for all parameter types of the two models in the different sample-size conditions are provided in Tables 3.9 – 3.12 in the appendix to Chapter 3.)

3.5.2.1 Root Median Standard Error

Figure 3.1 shows an effect of the sample size on the estimation bias of parameter types regarding the rmGPCM-3 (Figure 3.1a) and mPCM-3 (Figure 3.1b). For both models, the *RMdSE* values generally decreased by increasing the sample size with the exception of the class-specific variances of the latent trait variable that were accurately estimated already with the smallest sample size (maximal $RMdSE^{N=500}$ = .04 and .05 for the rmGPCM-3 and mPCM-3, respectively). Furthermore, the class-specific class-size

Table 3.3. Convergence rates of the EM algorithm and the Newton-Raphson algorithm, the number of required iterations, boundary values and improper solutions, and mean classification probabilities for the rmGPCM-3 and mPCM-3.

N	Conv. EM, %	Md_{EM} (Range$_{\text{EM}}$)	Conv. NR, %	Md_{NR} (Range$_{\text{NR}}$)	BV$_{\text{SE}}$, % (improper)	$M_{\text{P(Y\|G)}}$
rmGPCM-3						
500	100	303 (111 – 2133)	84	8 (5 – 600)	16 (3)	0.88
1000	100	256 (99 – 1256)	79	8 (4 - 600)	21 (1)	0.85
1500	100	197 (67 – 1467)	69	9 (3 - 600)	31(0)	0.84
2000	100	160 (58 – 1158)	68	9 (3 – 600)	32(0)	0.83
2500	100	139 (59 – 956)	72	9 (3 – 600)	28 (1)	0.82
3000	100	127 (54 – 402)	68	9 (3 – 600)	32 (1)	0.82
3500	100	122 (58 – 1897)	65	9 (3 – 600)	35(0)	0.82
4000	100	110 (49 – 377)	62	9 (3 – 600)	38(0)	0.82
4500	100	109 (51 – 373)	62	9 (3 - 600)	38(0)	0.82
5000	100	99 (50 – 603)	56	10 (3 – 600)	44 (1)	0.82
mPCM-3						
500	100	293 (126-1188)	96	8 (6-600)	4.4(0)	0.89
1000	100	247 (75-1415)	98	8 (4-600)	2.2(0)	0.86
1500	100	208 (70-1033)	99	8 (3-600)	0.8(0)	0.85
2000	100	171 (66-920)	99.6	8 (3-600)	0.4(0)	0.84
2500	100	151 (50-719)	99.8	7 (3-600)	0.2(0)	0.84
3000	100	138 (60-677)	100	7 (2-23)	0	0.84
3500	100	124 (57-497)	100	6 (2-19)	0	0.83
4000	100	119 (53-506)	100	6 (2-21)	0	0.83
4500	100	112 (47-541)	100	6 (2-14)	0	0.83
5000	100	110 (55-331)	100	5 (3-17)	0	0.83

Notes. *N*: sample-size condition. Conv.EM: convergence rate of the EM algorithm. Md_{EM} (Range$_{\text{EM}}$): median (range) of iterations required to reach a convergent solution of the EM algorithm. Conv.NR: the convergence rate of the Newton-Rapson algorithm. Md_{NR} (Range$_{\text{NR}}$): median (range) of iterations required to reach a convergent solution of the Newton-Rapson algorithm (Note, solutions with 600 iterations did not converge). BV$_{\text{SE}}$ (improper): the proportion of replications with boundary values (the number of replications with an improper solution). $M_{\text{P(Y|G)}}$: mean classification probability.

parameters were also only slightly biased (maximal $RMdSE^{N=500}$ = .23 and .22 and maximal $RMdSE^{N=5000}$ = .06 and .08 for the rmGPCM-3 and mPCM-3, respectively). In contrast, the estimation bias was higher for both types of item parameters across all sample-size conditions (for class-specific delta beta parameters, maximal $RMdSE^{N=500}$ = .76 and .80 and maximal $RMdSE^{N=5000}$ = .18 and .21 for the rmGPCM-3 and mPCM-3, respectively; for item discrimination parameters of the rmGPCM-3, maximal $RMdSE^{N=500}$ = .52 and maximal $RMdSE^{N=5000}$ = .16). For both types of item parameters, the *RMdSE* curves show an inflection point at *N* = 1,500, indicating a sufficient decline in bias up to this sample size while further increasing the sample size had only a slight effect on the reduction of the *RMdSE* values. Discrimination parameters of items possessing higher discrimination power were estimated less accurately (item 4). Furthermore, the class size additionally affected the amount of bias for the class-specific parameters (such as class sizes and delta beta parameters). In particular, the parameters of the largest class (g2) were estimated more accurately compared to those of the smaller classes (g1 and g3).

3.5.2.2 Standard Error Bias

The accuracy of standard error estimates for the rmGPCM-3 and mPCM-3 is illustrated in Figures 3.2a and 3.2b, respectively. In general, the results are almost identical to those reported for the *RMdSE*. A slight bias was found for the standard errors of latent variances (maximal $bias_{\mathrm{SE}}^{N=500}$ = .02 and .03 for the rmGPCM-3 and mPCM-3, respectively) and the class-size parameters (maximal $bias_{\mathrm{SE}}^{N=500}$ = .12 and .11 for the rmGPCM-3 and mPCM-3, respectively). In addition, the standard error estimates of item parameters were more biased in the case of small sample sizes but they showed a rapid reduction of bias by increasing the sample size: The bias was below .10 from *N* = 1,500 on for standard error estimates of the discrimination parameters and from *N* = 2,000 on for those of the delta beta parameters of both models. Exceptionally, the standard error bias of delta beta parameters of the small class (g3) could be accurately estimated from *N* = 2,500 and *N* = 3,000 on for the rmGPCM-3 and mPCM-3, respectively. In addition, item discrimination size and class sizes had an additional effect on standard error bias values.

3.5.2.3 Median Width of the Confidence Interval

Similar tendencies were also found for the width of confidence intervals for model parameter estimates (see Figures 3.3a and 3.3b for the rmGPCM-3 and mPCM-3, respectively). Small standard errors and consequently narrow confidence intervals were estimated primarily for both the latent variances (maximal $Md_{\mathrm{width_{CI}}}^{N=500}$ = .20 for both models) and class sizes (maximal $Md_{\mathrm{width_{CI}}}^{N=500}$ = .80 for both models) even with small sample size. Again, confidence intervals of item parameters were comparably wider (for delta beta parameters, maximal $Md_{\mathrm{width_{CI}}}^{N=500}$ = 3.15 and 3.40 and maximal $Md_{\mathrm{width_{CI}}}^{N=5000}$ = 1.04 and 1.15 for the

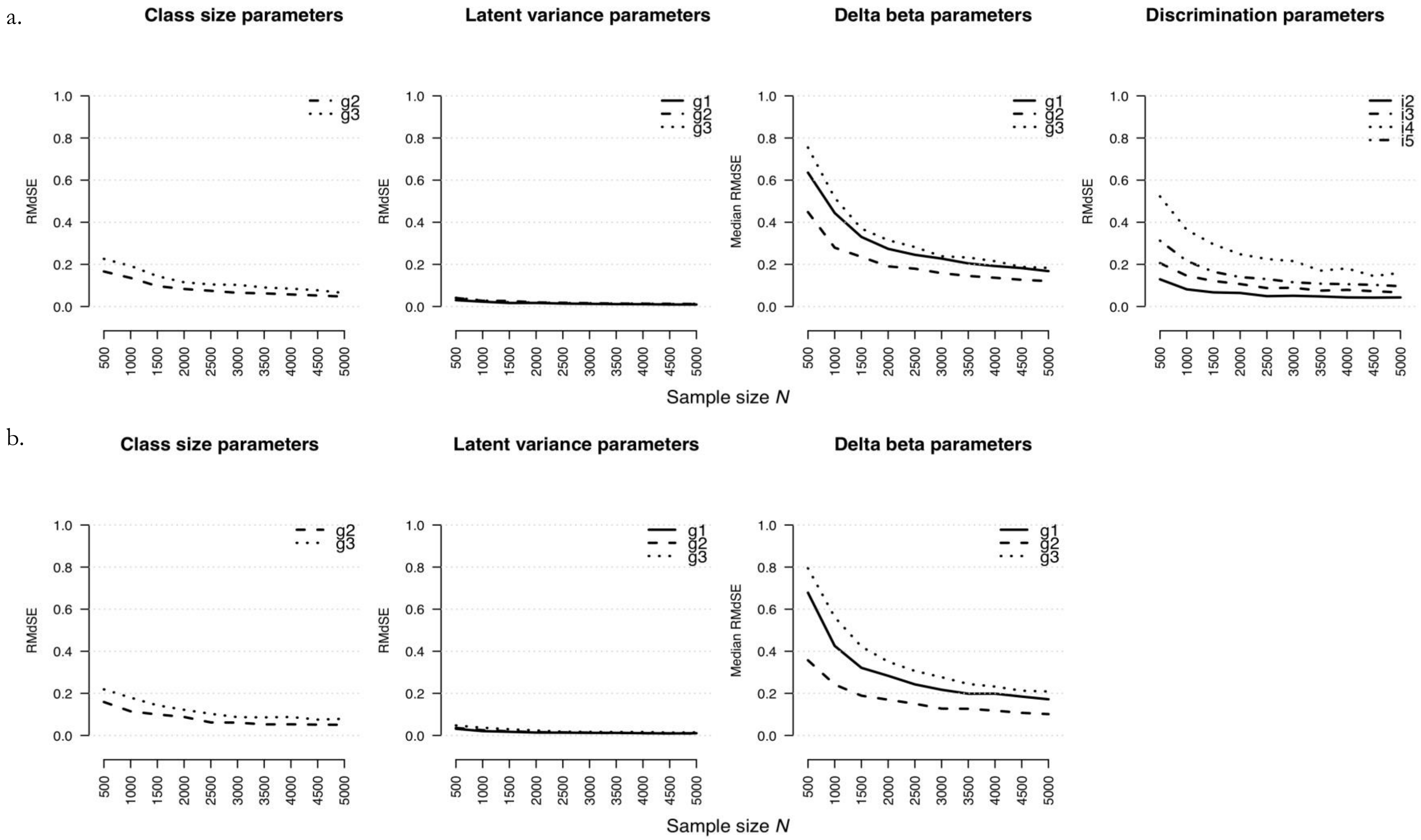
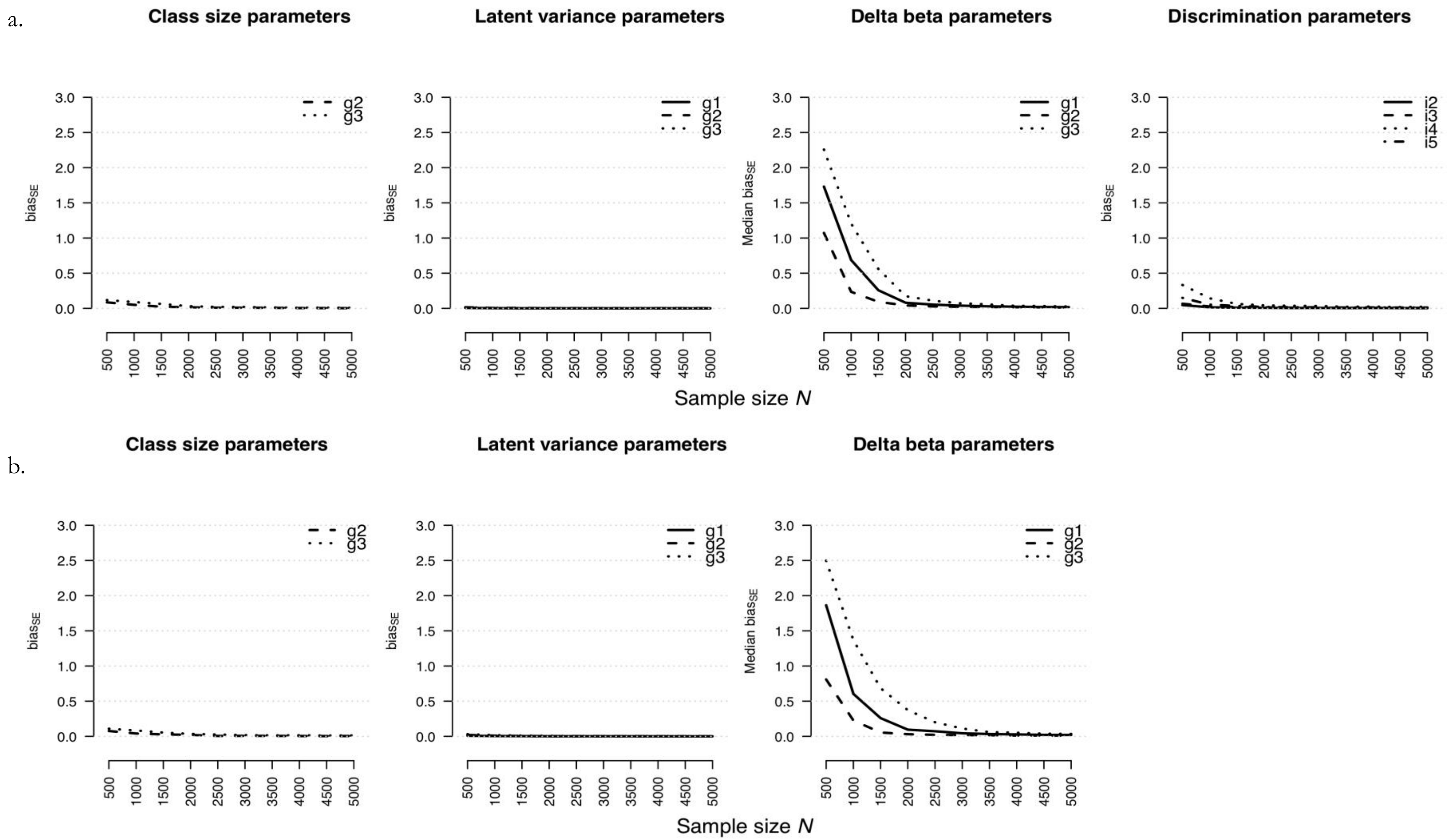


Figure 3.1. Root median squared error for parameter estimates in a) the rmGPCM-3 and b) mPCM-3.


a.

Class size parameters

Latent variance parameters

Delta beta parameters

Discrimination parameters

b.

Class size parameters

Latent variance parameters

Delta beta parameters

Figure 3.2. Bias of standard error estimates for a) the rmGPCM-3 and b) mPCM-3.

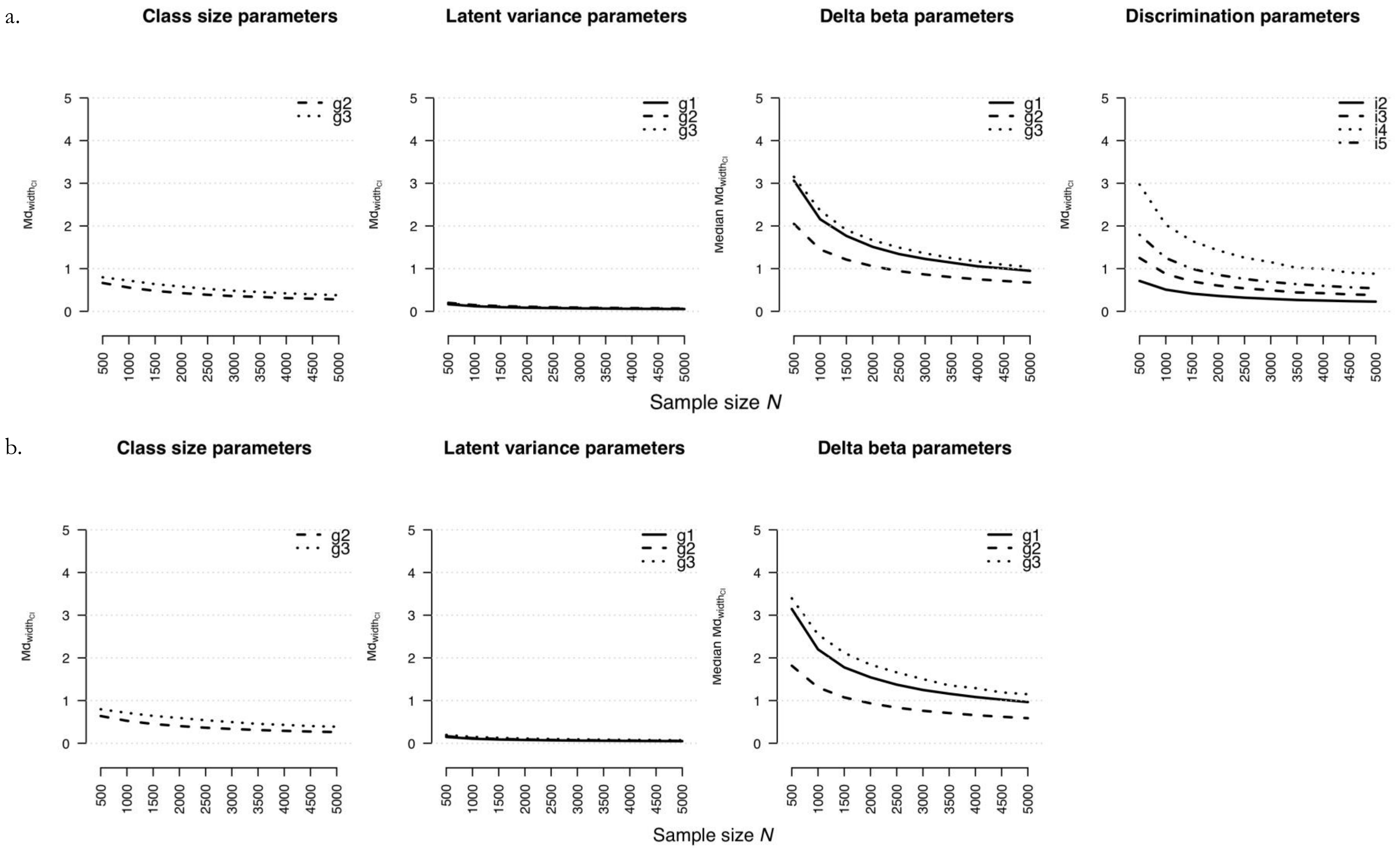


Figure 3.3. Width of the confidence interval for parameter estimates in a) the rmGPCM-3 and b) mPCM-3.

rmGPCM-3 and mPCM-3, respectively; for discrimination parameters, maximal $Md^{N=500}_{\mathrm{width}_{\mathrm{CI}}}$ = 2.97 and maximal $Md^{N=5000}_{\mathrm{width}_{\mathrm{CI}}}$ = .88 for the rmGPCM-3). For these parameters, the inflection point was observed at N = 1,500 with small confidence intervals from that point on. Identically, larger standard errors and consequently wider confidence intervals were also found for class-specific parameters of smaller classes and large discrimination parameters.

3.5.2.4 Estimation Accuracy for Specific Delta Beta Parameters

All three accuracy indices pointed out that, especially the delta beta parameters and their standard errors of both models were more biased compared to other parameter types. In particular, it concerned the first five delta beta parameters (see Figures 3.4, 3.6, 3.8 and 3.5, 3.7, 3.9 in the appendix to Chapter 3 for the rmGPCM-3 and mPCM-3, respectively). The first delta beta parameter in the DRS class (g2) showed high accuracy indices even with the largest sample size. Primarily, it may be caused by the low frequency of the lower categories expected for all classes, but especially for the DRS class (see Tables 3.13 and 3.14 in the appendix to Chapter 3 for the rmGPCM-3 and mPCM-3, respectively). By contrast, the upper five delta beta parameters were estimated more accurately already with the medium-sized samples. Furthermore, the amount of bias of the delta beta parameters is linked to a particular response style. The accuracy indices were smaller in the ERS and semi-ERS classes (g1 and g3, respectively) for the lower and the upper delta beta parameters and in DRS class for the middle ones.

3.5.2.5 Spearman's Rank Correlation

Table 3.4 reports averaged Spearman's rank correlations between generating and estimated delta beta parameters. In general, for both models, the correlation coefficients in latent classes increased with enlarging the sample size, indicating that the order of estimated parameters is more and more in accordance with that of the generating parameters. Delta beta parameters of the rmGPCM-3 showed a high concordance in order (above r_s = 0.90) with at least N = 1,500 observations for two first classes (ERS class and DRS class). For the small class (semi-ERS class), a larger sample was needed (at least N = 3,500), primarily because the delta beta parameters of this class were generally less accurately estimated (as reported above). For the mPCM-3, we found very similar results (see right column of Table 3.4).

Table 3.4. Averaged Spearman's rank correlations between the generating and estimated $\Delta\beta^{i}_{0sg}$-parameters for the rmGPCM-3 and mPCM-3.

	rmGPCM-3			**mPCM-3**		
N	**g1**	**g2**	**g3**	**g1**	**g2**	**g3**
500	.766	.703	.585	.762	.746	.576
1000	.875	.850	.707	.884	.874	.695
1500	.932	.913	.775	.935	.932	.730
2000	.950	.954	.833	.964	.959	.806
2500	.971	.971	.869	.977	.979	.847
3000	.980	.983	.891	.986	.986	.869
3500	.988	.988	.919	.996	.989	.897
4000	.991	.995	.920	.997	.993	.902
4500	.995	.995	.931	.999	.998	.923
5000	.996	.999	.930	.999	.997	.919

Note. g1, g2, and g3 indicate three latent classes of two models.

3.5.2.6 Coverage

Table 3.5 reports the coverage values for parameter types of the two generating models. In general, class-size parameters showed good coverage (≥ .90) from the medium-sized samples (*N* = 2,500) on. The class-specific variances of the latent trait demonstrated good coverage rate (≥ .90) even for the relatively small sample for the rmGPCM-3 (from *N* = 1,000 on) and with medium-sized samples for the mPCM-3 (from *N* = 2,500 on). In the case of small samples, the insufficient coverage of these parameter types can be explained by too narrow confidence intervals resp. small standard errors. In addition, item parameters generally achieved acceptably high coverage in all sample-size conditions (above .94 for discrimination parameters and above .93 and .94 for delta beta parameters of the rmGPCM-3 and mPCM-3, respectively).

To conclude from the results of this section, the two models mostly showed similar trends of estimation accuracy with varying sample size. Primarily, an accurate estimation of item parameters and their standard errors generally requires a larger sample (at least 1,500 – 2,000 observations) than the other parameter types. On the contrary, class-size parameters and variances of a latent trait could reach a high coverage rate with at least 2,500 observations. Beyond the sample size, both the size of the latent classes and the expected category frequencies are further influential factors for estimation accuracy.

Table 3.5. Coverage for parameters of the rmGPCM-3 and mPCM-3.

		N									
Parameter type	**Class**	**500**	**1000**	**1500**	**2000**	**2500**	**3000**	**3500**	**4000**	**4500**	**5000**
						rmGPCM-3					
π_g	2	**.82**	**.85**	.91	.91	.92	.93	.93	.94	.94	.95
	3	**.80**	**.81**	**.84**	**.89**	.92	.90	.92	.92	.93	.93
λ_g	1	.92	.92	.94	.95	.94	.94	.94	.92	.92	.95
	2	**.88**	.91	.91	.93	.94	.94	.96	.93	.94	.92
	3	.92	.90	.93	.90	.93	.91	.95	.91	.93	.91
$\Delta\beta^i_{0sg}$[1]	1	.96	.96	.96	.96	.96	.95	.96	.95	.96	.95
	2	.94	.93	.94	.94	.94	.94	.94	.94	.94	.94
	3	.95	.95	.96	.96	.96	.96	.95	.96	.95	.95
λ^i[1]		.95	.96	.96	.96	.95	.95	.95	.94	.94	.94
						mPCM-3					
π_g	2	**.82**	**.85**	**.88**	.90	.94	.92	.91	.95	.95	.93
	3	**.78**	**.83**	**.85**	**.86**	.91	.92	.92	.91	.90	.92
λ_g	1	**.88**	.91	.92	.93	.93	.91	.94	.92	.94	.90
	2	**.87**	**.89**	.91	.93	.94	.95	.94	.94	.95	.95
	3	**.84**	**.89**	**.88**	**.88**	.92	.94	.95	.91	.94	.91
$\Delta\beta^i_{0sg}$[1]	1	.97	.96	.96	.96	.96	.95	.96	.95	.95	.95
	2	.94	.94	.94	.94	.94	.94	.95	.95	.95	.95
	3	.95	.96	.96	.96	.96	.96	.96	.96	.96	.96

Notes. π_g: latent class-size parameter. λ_g: estimate of the variance of the class-specific latent trait variable. $\Delta\beta^i_{0sg}$: delta beta parameter. λ^i: item discrimination parameter.

[1] Mean coverage is reported for this parameter type.

Coverage rate under .90 is shown in bold.

3.5.3 Model Selection

Table 6 reports the proportion of replications in which the true class solution of the two population models was correctly identified as the best-fitting solution by the examined information criteria as well as their under- or overestimation rate across sample-size conditions. (Conditions with a proper performance are marked in bold.) For the rmGPCM, the AIC3 was the best criterion for selecting the three-class solution from medium-sized samples (from *N* = 1,500), followed by the SABIC (from *N* = 2,500). In contrast, the BIC and CAIC constantly underestimated the true number of classes in the conditions with small and medium-sized samples but they properly worked primarily with large samples (from *N* = 4,500 / 5,000, respectively). Whereas the AIC was showed a consistent tendency to overestimate the true number of classes (with only 79% - 84% success rate across sample-size conditions). Referring to the mPCM, the results are similar to those of the rmGPCM.

3.6 Discussion

The results of the present simulation study are useful for researchers interested in applying mixed polytomous IRT models for analyzing rating scales with many response categories that are widely used in the social and behavioral sciences. Because of the many response categories, it is likely to be confronted with the problem of sparse tables when different items are analyzed together. Therefore, the question of what sample size is required for the proper performance of the model is of high importance. Because only very few simulation studies have been conducted to examine mixed polytomous IRT models in general, and no simulation studies were found that considered the performance of these models under the data condition that is typically observed in survey studies (short scale with a large number of response categories), this application-oriented simulation study focused on the sample-size requirements for two models, the rmGPCM and mPCM, that are useful for exploring category use when applied to such data. Unlike most previous research, we took data-generating model parameters from an empirical model application in order to ensure ecological validity. We compared two models by manipulating only the sample size.

Study results indicated the effectivity of the EM algorithm to achieve a convergent solution for the mixed polytomous IRT models independently of model complexity and sample size. For more complex mixed IRT models as the rmGPCM-3 (as well as for overfitting models like the rmGPCM-4 or the mPCM-4), the NR method often produced non-convergent solutions in all sample-size conditions. In contrast, for the more parsimonious model (the mPCM-3), this problem occurred to a small extent and disappeared from the medium-sized sample (*N* = 3,000) on. Because of this failure of the NR method

Table 3.6. Proportion of replications with the best-fitting class solution identified by information criteria for the rmGPCM and mPCM.

	AIC				BIC				CAIC				AIC3				SABIC			
N	**g1**	**g2**	**g3[1]**	**g4**	**g1**	**g2**	**g3[1]**	**g4**	**g1**	**g2**	**g3[1]**	**g4**	**g1**	**g2**	**g3[1]**	**g4**	**g1**	**g2**	**g3[1]**	**g4**
										rmGPCM										
500	0	17	77	6	98	2	0	0	100	0	0	0	0	98	2	0	0	99	1	0
1000	0	0	84	16	1	99	0	0	17	83	0	0	0	37	63	0	0	91	9	0
1500	0	0	81	19	0	100	0	0	0	100	0	0	0	1	**99**	0	0	56	44	0
2000	0	0	80	20	0	100	0	0	0	100	0	0	0	0	**100**	0	0	16	84	0
2500	0	0	79	21	0	100	0	0	0	100	0	0	0	0	**100**	0	0	2	**98**	0
3000	0	0	82	18	0	92	8	0	0	100	0	0	0	0	**100**	0	0	0	**100**	0
3500	0	0	81	19	0	69	31	0	0	97	3	0	0	0	**100**	0	0	0	**100**	0
4000	0	0	78	22	0	21	79	0	0	66	34	0	0	0	**100**	0	0	0	**100**	0
4500	0	0	82	18	0	5	**95**	0	0	40	60	0	0	0	**100**	0	0	0	**100**	0
5000	0	0	80	20	0	0	**100**	0	0	7	93	0	0	0	**100**	0	0	0	**100**	0
										mPCM										
500	0	18	78	4	96	04	0	0	100	0	0	0	0	98	2	0	0	99	1	0
1000	0	0	86	14	0	100	0	0	13	87	0	0	0	33	67	0	0	87	13	0
1500	0	0	83	17	0	100	0	0	0	100	0	0	0	1	**99**	0	0	47	53	0
2000	0	0	82	18	0	100	0	0	0	100	0	0	0	0	**100**	0	0	12	88	0
2500	0	0	81	19	0	99	1	0	0	100	0	0	0	0	**100**	0	0	1	**99**	0
3000	0	0	80	20	0	85	15	0	0	100	0	0	0	0	**100**	0	0	0	**100**	0
3500	0	0	80	20	0	39	61	0	0	79	21	0	0	0	**100**	0	0	0	**100**	0
4000	0	0	79	21	0	13	87	0	0	53	47	0	0	0	**100**	0	0	0	**100**	0
4500	0	0	78	22	0	1	**99**	0	0	11	89	0	0	0	**100**	0	0	0	**100**	0
5000	0	0	82	18	0	0	**100**	0	0	3	**97**	0	0	0	**100**	0	0	0	**100**	0

Notes. AIC: Akaike's information criterion. BIC: Bayesian information criterion. CAIC: Consistent Akaike's Information Criterion. AIC3: Akaike Information Criterion. SABIC: sample-size adjusted BIC. g1, g2, g3 and g4 indicate a one-class to a four-class solution as a best-fitting model, respectively.

[1] The true class solution. A sufficient proportion of replications selected by a certain information criterion is shown in bold.

to work well in the context of complex models in the presence of sparse data, it could be recommended for researchers whose intention is to apply the rmGPCM or an other complex model to use only the EM algorithm (Vermunt & Magidson, 2013).

For the best-fitting three-class solution of the rmGPCM and mPCM, the accuracy of parameter and standard error estimates was evaluated. For both models, all parameter types (class-size parameters, class-specific variances of the latent trait variable, class-specific delta beta parameters, and item discrimination parameters only in the rmGPCM) and their corresponding standard errors mostly indicated the same trends. First, the estimation accuracy of parameters and standard errors improved as sample size increased. Specifically, delta beta parameters mainly showed slight improvement and appropriately reproduced true order within items (r_s > .90) from the sample size of 1,500 observations (with the exception of the small class g3 concerning the last point). Precise standard errors ($bias_{SE}$ < .10) could be obtained only from N = 2,000 on (with the same exception of the class g3). Similar results were observed for discrimination parameters and their standard errors. In turn, class-size parameters and class-specific variances of the latent trait and corresponding standard errors were estimated pretty accurately even with small sample sizes. To obtain appropriate coverage rates for these parameters at least 2,500 observations were however necessary. That may be explained by narrow confidence intervals due to small standard errors for these parameters compared to those of item parameters. Second, we observed that class-specific parameters and their standard errors are more precisely estimated in the largest class (g2) and less accurately in the small class (g3). For example, we found that for estimating delta beta parameters of the small class appropriately and to reproduce their true order of the population model sample sizes of 3,500 and 4,000 observations are necessary for the rmGPCM-3 and mPCM-3, respectively. An effect of the class size on estimation accuracy has been already pointed out in previous research (Cho, 2014; Preinerstorfer & Formann, 2012). Third, class-specific delta beta parameters and standard errors of the categories preferred in latent classes were estimated more accurately. For example, the first delta beta parameter and its standard error, especially in the ORS class (g2), was extremely biased due to very low expected frequencies of the lower categories. By increasing the sample size, the bias was found to be partly compensated in the semi-ERS class (g3) but hardly in the ORS class (g2). The crucial relevance of sufficient category frequencies to gain satisfactory estimation accuracy and to avoid boundary and extreme values has been emphasized in previous research on traditional polytomous IRT models (DeMars, 2003; He & Wheadon, 2013). Fourth, discrimination parameters and standard errors of highly discriminating items were more strongly biased.

We conclude from our results that an application of both models with an assumed three-class mixture on short-scale data assessed with many response categories can be reasonable with the sample size of at least 2,500 observations. Compared to bias statistics from previous research, the estimation accuracy primarily of delta beta parameters of both models in this simulation study was somewhat lower.

However, in contrast to other simulation studies, the present study is based on empirically found parameters as true model parameters, which include unordered thresholds, nearly located parameters on the latent continuum, and some extreme parameters, as it is often the case in the real research studies. Moreover, due to the rating scale with many response categories, both models include many delta beta parameters within an item to be estimated. These specifics make the present simulation study unique and its results relevant for applied research. Nevertheless, researchers should be aware of the problem of low category frequencies that will probably occur in the context of the considered data situation and cause estimation problems (in form of boundary, extreme, and inaccurate parameter estimates) that may hardly be remedied only by increasing the sample size. Also, we discourage practitioners to use a small sample of fewer than 1,500 observations under which both mixture polytomous IRT models were especially unsuccessful in providing less biased estimates.

The last focus of this work was to examine five information criteria concerning their effectiveness to detect the true class solution of the rmGPCM and mPCM: the AIC, BIC, CAIC, AIC3, and SABIC. For both models, the best result was found for the AIC3 (99% accuracy of *N* = 1,500), following by the SABIC (98% accuracy of *N* = 2,500). This is consistent with the research in the field of finite mixture modeling, reporting that these information criteria are effective for identifying complex latent mixtures (above two classes) with sufficiently small sample sizes (Fonseca, 2010; Choi et al., 2017; Yu & Park, 2014). But these results are opposed to the research evidence which suggests that the BIC and CAIC are favorites for model selection in the context of mixed IRT models (e.g., Cho, 2014; Li et al., 2009). The present simulation study indicated that both these information criteria generally underestimated the true number of classes and worked well only for large samples (of *N* = 4,500/5,000, respectively). The BIC was superior to the CAIC, and the CAIC performed better for the one-parameter mixed IRT model (the mPCM). An explanation could be that in contrast to previous simulation studies, these information criteria had a very severe penalty term in the present study because of a large number of model parameters due to 11-point items apart from the sample size. In general, the results of this study showed that in the context of mixed polytomous IRT models the more effective information criteria are those that do not or slightly penalize the sample size used for the model application. In line with previous research in the context of mixed dichotomous IRT models (e.g., Cho et al., 2013), the AIC selected the correct class solution on average only in 80% of all cases and otherwise preferred an overparameterized model solution. Based on our results, we recommend that researchers use the AIC3 for the selecting the best-fitting solution of the rmGPCM and mPCM when data are assessed with many response categories and the sample size is medium-large (at least *N* = 1,500) and SABIC from *N* = 2,500 observations. In contrast, the BIC and CAIC performed well only with a large sample (of *N* = 4,500 / 5,000, respectively).

3.6.1 Limitations and Future Research

The generalization of the reported simulation results is limited due to the specificity of the data situation considered (short tests and long rating scales) and the latent mixture (three unequally-sized classes with specific patterns of category use). In addition, we did not include further useful mixed IRT models (e.g., the mixed NRM, the mixed GPCM with a random response-style effect or mixed multidimensional IRT models) because of their high complexity. Future research should expand the range of data conditions, latent mixtures, and mixed polytomous IRT models in accordance with further application fields of mixed IRT approach.

In general, simulation results depend on the used estimation method. In the present simulation study, model parameters were estimation by means of the maximum likelihood estimation method, which is implemented in Latent GOLD. However, more and more recent studies on mixed IRT models use the Bayesian estimation method, which is also flexible for model restrictions and extension. Therefore, future research should focus on the comparison of two estimation methods, as previous studies on mixed dichotomous IRT models has indicated benefits of the latter method with regard to parameter estimation bias and classification accuracy for short scales, small sample sizes, and complex latent mixtures (e.g., Finch & French, 2012).

Future studies should examine whether including external covariates into mixed polytomous IRT models may improve correct identification of underlying structure and accuracy of class assignment and parameter estimates, especially when data conditions or latent mixtures are challenging. Empirical evidence on this issue has been shown in the context of mixed dichotomous IRT models (see Dai, 2013; De la Torre & Hong, 2010; Smit, Kelderman, & van der Flier, 2000) but not yet for mixed polytomous IRT models.

3.6.2 Conclusion

The current application-oriented simulation study was aimed at identifying required sample size for the mixed one- and two parameter IRT models for polytomous data (rmGPCM, mPCM) and investigating diverse information criteria concerning their capacity to correctly detect the best-fitting model solution. Focusing on specific data situation present in panel surveys by assessing life satisfaction with a short scale and many response categories and on the latent mixture of typical category use patterns in that context, this simulation study produced results suggesting that two models exhibited similar trends of estimation accuracy at manipulated sample sizes. The sample size of fewer than 1,500 respondents was insufficiently small, and a sample size of 2,500 respondents seems to be sufficient. A further increase of the sample size had a positive effect on the estimation accuracy, especially in the small class, but was hardly helpful for extremely biased item parameters and standard errors arising in the case of low-frequency categories.

In particular, the mixed two-parameter IRT model (rmGPCM) indicated more estimation problems (in form of non-convergence of the Newton-Raphson algorithm, occurrence of extreme parameter estimates, and boundary standard error estimates) due to insufficient responses of a few categories as the mixed one-parametric IRT model did. Of information criteria, the AIC3, followed by the SABIC performed better as opposed to the BIC and CAIC for the recommended sample size.

3.7 References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716-723. doi: 10.1109/TAC.1974.1100705

Alexeev, N., Templin, J., & Cohen, A. S. (2011). Spurious latent classes in the mixture Rasch model. *Journal of Educational Measurement*, *48*, 313-332. doi: 10.1111/j.1745-3984.2011.00146.x

Andrews, R. L., & Currim, I. S. (2003). A comparison of segment retention criteria for finite mixture logit models. *Journal of Marketing Research, 40*, 235-243. doi: 10.1509/jmkr.40.2.235.19225

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*, 1235-1245. doi: 10.1016/j.paid.2005.10.018

Baghaei, P., & Carstensen, C. H. (2013). Fitting the Mixed Rasch model to a reading comprehension test: Identifying reader types. *Practical Assessment, Research & Evaluation, 18*, 1-13. Available online: http://pareonline.net/getvn.asp?v=18&n=5

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 29-51. doi: 10.1007/BF02291411

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*, 443-459. doi: 10.1007/BF02293801

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52*, 345-370. doi: 10.1007/BF02294361

Bozdogan, H. (1994). Mixture-model cluster analysis using model selection criteria and a new informational measure of complexity. In H. Bozdogan (Eds.), *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.

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*, 205-215. doi: 10.1177/014662169401800302

Cho, S. J., Cohen, A. S., & Kim, S. H. (2013). Markov chain Monte Carlo estimation of a mixture item response theory model. *Journal of Statistical Computation and Simulation*, *83*, 278-306. doi: 10.1080/00949655.2011.603090

Cho, S. J., Suh, Y., & Lee, W. Y. (2016). An NCME Instructional module on latent DIF analysis using mixture Item Response models. *Educational Measurement: Issues and Practice*, *35*, 48-61. https://doi.org/10.1111/emip.12093

Cho, Y. (2014). The mixture distribution polytomous rasch model used to account for response styles on rating scales: A simulation study of parameter recovery and classification accuracy. *Dissertation Abstracts International*, *75*. http://hdl.handle.net/1903/14511

Choi, I. H., Paek, I., & Cho, S. J. (2017). The impact of various class-distinction features on model selection in the mixture Rasch model. *The Journal of Experimental Education, 85,* 411-424. doi: 10.1080/00220973.2016.1250208

Choi, S. W., Cook, K. F., & Dodd, B. G. (1997). Parameter recovery for the partial credit model using MULTILOG. *Journal of Outcome Measurement, 1*, 114-142.

Dai, Y. (2013). A mixture Rasch model with a covariate: A simulation study via Bayesian Markov chain Monte Carlo estimation. *Applied Psychological Measurement*, *37*, 375-396. doi: 10.1177/0146621612475076

De Ayala, R. J., & Sava-Bolesta, M. (1999). Item parameter recovery for the nominal response model. *Applied psychological measurement*, *23*, 3-19. doi: 10.1177/01466219922031130

De la Torre, J., & Hong, Y. (2010). Parameter estimation with small sample size a higher-order IRT model approach. *Applied Psychological Measurement*, *34*, 267-285. doi: 10.1177/0146621608329501

De La Torre, J., Stark, S., & Chernyshenko, O. S. (2006). Markov chain Monte Carlo estimation of item parameters for the generalized graded unfolding model. *Applied Psychological Measurement*, *30*, 216-232. doi: 10.1177/0146621605282772

DeMars, C. E. (2003). Sample size and the recovery of nominal response model item parameters. *Applied Psychological Measurement*, *27*, 275-288. doi: 10.1177/0146621603027004003

Dias J. G. (2006). Latent class analysis and model selection. In M. Spiliopoulou, R. Kruse, C. Borgelt, A. Nürnberger, W. Gaul (Eds.), *From Data and Information Analysis to Knowledge Engineering. Studies in Classification, Data Analysis, and Knowledge Organization* (pp 95-102). Springer, Berlin, Heidelberg. doi: 10.1007/3-540-31314-1_10

Egberink, I. J., Meijer, R. R., & Veldkamp, B. P. (2010). Conscientiousness in the workplace: Applying mixture IRT to investigate scalability and predictive validity. *Journal of Research in Personality*, *44*, 232-244. doi: 10.1016/j.jrp.2010.01.007

Eid, M., & Rauber, M. (2000). Detecting measurement invariance in organizational surveys. *European Journal of Psychological Assessment, 16*, 20-30. doi: 10.1027//1015-5759.16.1.20

Eid, M., & Zuckar, M. (2007). Detecting response styles and faking in personality and organizational assessments by Mixed Rasch Models. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models* (pp. 255-270). New York: Springer Science + Business Media. doi: 10.1007/978-0-387-49839-3_16

Embretson, S. E., & Reise, S. P. (2013). Item response theory for psychologists. Hillsdale, New Jersey: Erlbaum.

Finch, W. H., & French, B. F. (2012). Parameter estimation with mixture item response theory models: A Monte Carlo comparison of maximum likelihood and Bayesian methods. *Journal of Modern Applied Statistical Methods*, *11*, 167-178. doi: 10.22237/jmasm/1335845580

Finch, W. H., & Pierson, E. E. (2011). A mixture IRT analysis of risky youth behavior. *Frontiers in psychology*, *2*. doi: 10.3389/fpsyg.2011.00098

Fonseca, J.R.S. (2010). On the performance of information criteria in latent segment models estimation with categorical segmentation base variables. In *Proceedings of ICMSE 2010. International Conference on Mathematical Science and Engineering, World Academy of Science, Engineering and Technology, WASET* (pp 330-337). Rio de Janeiro, Brazil.

French, G., & Dodd, B. (1999). Parameter recovery for the rating scale model using PARSCALE. *Journal of Outcome Measurement, 3*, 176–199.

Frick, H., Strobl, C., & Zeileis, A. (2015). Rasch mixture models for DIF detection: A comparison of old and new score specifications. *Educational and Psychological Measurement*, *75*, 208-234. doi: 10.1177/0013164414536183

Gnaldi, M., Bacci, S., & Bartolucci, F. (2016). A multilevel finite mixture item response model to cluster examinees and schools. *Advances in Data Analysis and Classification*, *10*, 53-70. doi: 10.1007/s11634-014-0196-0

He, Q., & Wheadon, C. (2013). The effect of sample size on item parameter estimation for the partial credit model. *International Journal of Quantitative Research in Education*, *1*, 297-315. doi: 10.1504/IJQRE.2013.057692

Huang, H. Y. (2016). Mixture random-effect IRT models for controlling extreme response style on rating scales. *Frontiers in Psychology*, *7*. doi: 10.3389/fpsyg.2016.01706

Jasper, F., Nater, U. M., Hiller, W., Ehlert, U., Fischer, S., & Witthöft, M. (2013). Rasch scalability of the somatosensory amplification scale: a mixture distribution approach. *Journal of Psychosomatic Research*, *74*, 469-478. doi: 10.1016/j.jpsychores.2013.02.006.

Jensuttiwetchakul, P., Kanjanawasee, S., & Ngudgratoke, S. (2016). A development of the 3PL MMM-IRT model for identifying latent class. *Procedia-Social and Behavioral Sciences*, *217*, 719-728. doi: 10.1016/j.sbspro.2016.02.132

Jin, K. Y., & Wang, W. C. (2014). Item response theory models for performance decline during testing. *Journal of Educational Measurement*, *51*, 178-200. doi: 10.1111/jedm.12041

Kieftenbeld, V., & Natesan, P. (2012). Recovery of graded response model parameters: A comparison of marginal maximum likelihood and Markov chain Monte Carlo estimation. *Applied Psychological Measurement*, *36*, 399-419. doi: 10.1177/0146621612446170

Krosnick, J.A., & Presser, S. (2010). Question and questionnaire design. In P.V. Marsden, J.D. Wright (Eds.), *Handbook of survey research* (pp. 263-314). Bingley, UK: Emerald Group Publishing.

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*. doi: 10.3389/fpsyg.2016.01998

Lange, R. (2008). Binary items and beyond: A simulation of computer adaptive testing using the Rasch partial credit model. *Journal of Applied Measurement, 9*, 81–104. Retrieved from http://search.proquest.com/docview/622124039?accountid=14521

Li, F., Cohen, A. S., Kim, S. H., & Cho, S. J. (2009). Model selection methods for mixture dichotomous IRT models. *Applied Psychological Measurement, 33*, 353-373. doi: 10.1177/0146621608326422

Maij-de Meij, A. M., Kelderman, H., & van der Flier, H. (2008). Fitting a mixture item response theory model to personality questionnaire data: Characterizing latent classes and investigating possibilities for improving prediction. *Applied Psychological Measurement, 32*, 611-631. doi: 10.1177/0146621607312613

Meiser, T., & Machunsky, M. (2008). The personal structure of personal need for structure. *European Journal of Psychological Assessment*, *24*, 27-34. doi: 10.1027/1015-5759.24.1.27

Meyer, J. P., & Hailey, E. (2012). A study of Rasch, partial credit, and rating scale model parameter recovery in WINSTEPS and jMetrik. *Journal of Applied Measurement, 13*, 248–258.

Mneimneh, Z. N., Heeringa, S. G., Tourangeau, R., & Elliott, M. R. (2014). Bridging psychometrics and survey methodology: Can mixed Rasch models identify socially desirable reporting behavior?. *Journal of Survey Statistics and Methodology*, *2*(3), 257-282. doi:1093/jssam/smu008

Muraki, E. (1997). A generalized partial credit model. In W. van der Linden & R. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 153-164). New York: Springer.

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*, 535-569. doi: 10.1080/10705510701575396

Preinerstorfer, D., & Formann, A. K. (2012). Parameter recovery and model selection in mixed Rasch models. *British Journal of Mathematical and Statistical Psychology*, *65*, 251-262. doi: 10.1111/j.2044-8317.2011.02020.x

Reise, S. P., & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of educational Measurement*, *27*, 133-144. doi: 10.1111/j.1745-3984.1990.tb00738.x

Rost, J. (1997). Logistic mixture models. In W. van der Linden & R. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 449-463). New York: Springer.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461-464. Available online: http://www.jstor.org/stable/2958889

Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika, 52*, 333-343. doi: 10.1007/BF02294360

Smit, A., Kelderman, H., & van der Flier, H. (2000). The mixed Birnbaum model: Estimation using collateral information. *Methods of Psychological Research Online, 5*, 31-43.

Smith, E. V. Jr., 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*, 23-40.

Summerfield, M., Freidin, S., Hahn, M., Li, N., Macalalad, N., Mundy, L., et al. (2015). *HILDA User Manual – Release 14*. Melbourne, VIC: Melbourne Institute of Applied Economic and Social Research, University of Melbourne.

Tietjens, M., Freund, P. A., Büsch, D., & Strauss, B. (2012). Using mixture distribution models to test the construct validity of the Physical Self-Description Questionnaire. *Psychology of Sport and Exercise*, *13*, 598-605. doi: 10.1016/j.psychsport.2012.02.009

van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. Springer New York.

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*, 195-217. doi: 10.1093/ijpor/eds021

Vermunt, J. K., & Magidson, J. (2006). *Latent GOLD 4.0 and IRT modeling.* Belmont, MA: Statistical Innovations Inc.

Vermunt, J.K., & Magidson J. (2013). *Technical Guide for Latent GOLD 5.0: Basic, Advanced, and Syntax.* Belmont, MA: Statistical Innovations Inc.

Vermunt, J.K., & Magidson, J. (2008). *LG-Syntax User's Guide: Manual for Latent GOLD 4.5 Syntax Module.* Belmont, MA: Statistical Innovations Inc.

Vermunt, J.K., Magidson, J. (2016). *Technical Guide for Latent Gold 5.1: Basic, Advanced, and Syntax*. Belmont, MA: Statistical Innovations, Inc.

von Davier, M., & Carstensen, C. H. (2007). *Multivariate and mixture distribution Rasch models*. New York: Springer.

von Davier, M., & Yamamoto, K. (2004). Partially observed mixtures of IRT models: An extension of the generalized partial-credit model. *Applied Psychological Measurement, 28*, 389-406. doi: 10.1177/0146621604268734

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., Kuhn, U., Lipps, O., Ryser, V. A., ... & Wernli, B. (2014). *Swiss household panel user guide (1999-2013).* Lausanne: Swiss Household Panel.

Wagner-Menghin, M. M. (2006). The mixed-Rasch model: An example for analyzing the meaning of response latencies in a Personality Questionnaire. *Journal of Applied Measurement*, *7*, 225-237.

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*, 956-972. doi: 10.1177/0013164404268674

Wetzel, E., Böhnke, J. R., & Rose, N. (2016). A simulation study on methods of correcting for the effects of extreme response style. *Educational and Psychological Measurement, 76*(2), 304-324. doi: 10.1177/0013164415591848

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*, 178-189. doi: 10.1016/j.jrp.2012.10.010

Willits, F. K., Theodori, G. L., & Luloff, A. E. (2016). Another look at Likert scales. *Journal of Rural Social Sciences*, *31*(3), 126-139.

Wu, P. C., & Huang, T. W. (2010). Person heterogeneity of the BDI-II-C and its effects on dimensionality and construct validity: using mixture item response models. *Measurement and Evaluation in Counseling and Development*, *43*, 155-167. doi: 10.1177/0748175610384808

Yang, C. C., & Yang, C. C. (2007). Separating latent classes by information criteria. *Journal of Classification, 24*, 183-203. doi: 10.1007/s00357-007-0010-1

Yu, H. T., & Park, J. (2014). Simultaneous decision on the number of latent clusters and classes for multilevel latent class models. *Multivariate behavioral research, 49*, 232-244. doi: 10.1080/00273171.2014.900431

Ziegler, M., & Kemper, C. J. (2013). Extreme response style and faking: Two sides of the same coin?. In P. Winker, N. Menold, & R. Porst (Eds.), *Interviewers deviations in surveys – impact, reasons, detection and prevention* (pp. 217–233). Frankfurt a.M.: Peter Lang.

3.8 Appendix to Chapter 3

Table 3.7. Convergence rates of the EM algorithm and the Newton-Raphson algorithm, the number of required iterations, boundary values and improper solutions, and mean classification probabilities for the rmGPCM.

N	**Conv. EM, %**	***Md*$_{\text{EM}}$ (Range$_{\text{EM}}$)**	**Conv. NR, %**	***Md*$_{\text{NR}}$ (Range$_{\text{NR}}$)**	**BV$_{\text{SE}}$, % (improper)**	***M*$_{\text{P(Y\|G)}}$**
			rmGPCM-1			
500	100	25 (2-148)	100	4 (1-16)	0	1.00
1000	100	25 (14-61)	100	4 (3-16)	0	1.00
1500	100	22 (11-76)	100	4 (2-12)	0	1.00
2000	100	21 (12-57)	100	4 (2-11)	0	1.00
2500	100	22 (12-59)	100	4 (2-10)	0	1.00
3000	100	22 (13-51)	100	4 (3-12)	0	1.00
3500	100	21 (2-50)	100	4 (2-10)	0	1.00
4000	100	22 (13-53)	100	4 (2-9)	0	1.00
4500	100	20 (13-57)	100	4 (2-13)	0	1.00
5000	100	21 (14-77)	100	4 (2-11)	0	1.00
			rmGPCM-2			
500	100	133 (47-525)	99.6	8 (3-600)	0.4 (0)	0.93
1000	100	88 (31-501)	99.8	8 (3-600)	0.2 (0)	0.92
1500	100	67 (34-380)	100	5 (2-79)	0	0.91
2000	100	60 (30-377)	100	4 (3-43)	0	0.91
2500	100	56 (32-180)	100	4 (2-26)	0	0.91
3000	100	55 (30-231)	100	4 (2-17)	0	0.91
3500	100	58 (28-250)	100	4 (2-40)	0	0.91
4000	100	51 (27-227)	100	4 (2-33)	0	0.91
4500	100	56 (30-174)	100	3 (2-37)	0	0.91
5000	100	51 (25-118)	100	4 (2-39)	0	0.91
			rmGPCM-4			
500	100	637 (274-5396)	50.2	591 (6-600)	49.8 (9)	0.87
1000	100	685 (291-2536)	35.0	600 (6-600)	64.8 (5)	0.82
1500	100	788 (296-2931)	26.0	600 (7-600)	73.8 (14)	0.80
2000	100	856 (305-3803)	22.6	600 (7-600)	77.2 (15)	0.79
2500	100	833 (258-3975)	22.8	600 (7-600)	77.0 (10)	0.78
3000	100	915 (310-3463)	22.8	600 (7-600)	77.2 (12)	0.77
3500	100	923 (346-3478)	15.8	600 (7-600)	83.8 (15)	0.77
4000	100	931 (265-3151)	17.0	600 (5-600)	83.0 (14)	0.77
4500	100	938 (328-5614)	15.0	600 (6-600)	85.0 (6)	0.76
5000	100	952 (316-3744)	12.6	600 (7-600)	87.4 (10)	0.76

Notes. N: sample-size condition. Conv.EM: convergence rate of the EM algorithm. *Md*$_{\text{EM}}$ (Range$_{\text{EM}}$): median (range) of iterations required to reach a convergent solution of the EM algorithm. Conv.NR: the convergence rate of the Newton-Rapson algorithm. *Md*$_{\text{NR}}$ (Range$_{\text{NR}}$): median (range) of iterations required to reach a convergent solution of the Newton-Rapson algorithm (Note, solutions with 600 iterations did not converge). BV$_{\text{SE}}$ (improper): the proportion of replications with boundary values (the number of replications with an improper solution). *M*$_{\text{P(Y|G)}}$: mean classification probability.

Table 3.8. Convergence rates of the EM algorithm and the Newton-Raphson algorithm, the number of required iterations, boundary values and improper solutions, and mean classification probability for the mPCM.

N	**Conv. EM, %**	***Md*$_{\text{EM}}$ (Range$_{\text{EM}}$)**	**Conv. NR, %**	***Md*$_{\text{NR}}$ (Range$_{\text{NR}}$)**	**BV$_{\text{SE}}$, % (improper)**	***M*$_{\text{P(Y\|G)}}$**
			mPCM-1			
500	100	16 (7-37)	100	3 (2-9)	0	1.00
1000	100	15 (5-28)	100	3 (2-9)	0	1.00
1500	100	15 (6-33)	100	3 (2-7)	0	1.00
2000	100	14 (5-28)	100	3 (2-8)	0	1.00
2500	100	14 (7-26)	100	3 (2-8)	0	1.00
3000	100	14 (7-31)	100	3 (2-7)	0	1.00
3500	100	14 (7-32)	100	3 (2-9)	0	1.00
4000	100	13 (7-28)	100	3 (2-8)	0	1.00
4500	100	13 (7-32)	100	3 (2-9)	0	1.00
5000	100	13 (7-35)	100	3 (2-7)	0	1.00
			mPCM-2			
500	100	129 (44-666)	99.2	7 (4-600)	0.8 (0)	0.93
1000	100	84 (27-609)	99.8	8 (2-600)	0.2 (0)	0.92
1500	100	67 (29-586)	100	5 (2.18)	0	0.92
2000	100	56 (25-437)	100	4 (2-14)	0	0.92
2500	100	53 (30-174)	100	4 (3-15)	0	0.92
3000	100	52 (28-342)	100	4 (2-18)	0	0.91
3500	100	51 (28-216)	100	4 (2-19)	0	0.92
4000	100	48 (26-140)	100	4 (2-20)	0	0.91
4500	100	51 (27-132)	100	4 (2-32)	0	0.91
5000	100	47 (26-169)	100	3 (2-21)	0	0.91
			mPCM-4			
500	100	493 (196-2370)	71.4	10 (6-600)	28.4 (1)	0.87
1000	100	609 (234-2500)	64.0	12 (6-600)	36.0 (2)	0.83
1500	100	609 (185-4861)	66.6	10 (5-600)	33.0 (0)	0.81
2000	100	716 (244-2750)	67.2	10 (6-600)	32.8 (0)	0.81
2500	100	729 (277-4202)	67.2	9 (6-600)	32.8 (4)	0.80
3000	100	763 (216-3899)	64.6	10 (5-600)	35.2 (5)	0.79
3500	100	858 (287-4013)	62.8	13 (4-600)	37.2 (3)	0.79
4000	100	889 (253-5391)	69.6	10 (3-600)	30.2 (2)	0.79
4500	100	928 (259-6080)	68.0	10 (4-600)	32.0 (0)	0.79
5000	100	980 (269-3736)	65.8	10 (4-600)	34.0 (3)	0.79

Notes. N: sample-size condition. Conv.EM: convergence rate of the EM algorithm. *Md*$_{\text{EM}}$ (Range$_{\text{EM}}$): median (range) of iterations required to reach a convergent solution of the EM algorithm. Conv.NR: the convergence rate of the Newton-Rapson algorithm. *Md*$_{\text{NR}}$ (Range$_{\text{NR}}$): median (range) of iterations required to reach a convergent solution of the Newton-Rapson algorithm (Note, solutions with 600 iterations did not converge). BV$_{\text{SE}}$ (improper): the proportion of replications with boundary values (the number of replications with an improper solution). *M*$_{\text{P(Y|G)}}$: mean classification probability.

Table 3.9. Root median squared error for parameter estimates (*RMdSE*).

	rmGPCM-3						mPCM-3		
N	**Item**	λ^i	**Class**	$\Delta\beta_{0sg}$[1]	λ_g	π_g	$\Delta\beta_{0sg}$[1]	λ_g	π_g
500	2	.129	1	.636	.030		.678	.032	
	3	.206	2	.448	.041	.166	.357	.036	.159
	4	.523	3	.755	.041	.226	.795	.048	.219
	5	.312							
1000	2	.082	1	.444	.023		.426	.021	
	3	.146	2	.279	.027	.136	.242	.025	.115
	4	.364	3	.517	.029	.193	.562	.037	.180
	5	.216							
1500	2	.067	1	.330	.017		.321	.017	
	3	.121	2	.236	.026	.097	.189	.020	.100
	4	.295	3	.370	.020	.144	.422	.029	.143
	5	.165							
2000	2	.064	1	.274	.017		.283	.014	
	3	.107	2	.191	.019	.083	.170	.016	.088
	4	.248	3	.315	.019	.115	.351	.024	.122
	5	.140							
2500	2	.049	1	.245	.014		.243	.013	
	3	.088	2	.180	.018	.075	.151	.015	.062
	4	.225	3	.282	.016	.105	.306	.018	.103
	5	.130							
3000	2	.051	1	.228	.013		.217	.012	
	3	.089	2	.158	.016	.065	.128	.014	.061
	4	.216	3	.239	.016	.103	.277	.017	.087
	5	.115							
3500	2	.048	1	.205	.011		.198	.012	
	3	.075	2	.145	.014	.062	.127	.012	.052
	4	.170	3	.233	.014	.090	.245	.016	.086
	5	.109							
4000	2	.043	1	.192	.011		.199	.011	
	3	.079	2	.136	.013	.057	.118	.011	.053
	4	.181	3	.215	.014	.085	.232	.016	.088
	5	.106							
4500	2	.042	1	.182	.010		.185	.010	
	3	.073	2	.127	.013	.052	.107	.010	.051
	4	.145	3	.188	.012	.077	.212	.014	.075
	5	.103							
5000	2	.043	1	.168	.009		.172	.010	
	3	.067	2	.121	.012	.047	.101	.010	.051
	4	.160	3	.183	.013	.064	.209	.015	.079
	5	.096							

Note. [1] Median *RMdSE* for class-specific delta beta parameter estimates.

Table 3.10. Empirical standard deviation of parameter estimates ($SD_{\hat{p}}$).

	rmGPCM-3						mPCM-3		
N	Item	λ^i	Class	$\Delta\beta_{0sg}$[1]	λ_g	π_g	$\Delta\beta_{0sg}$[1]	λ_g	π_g
500	2	.219	1	2.348	.047		2.527	.046	
	3	.353	2	1.663	.067	.255	1.276	.063	.239
	4	1.024	3	2.855	.060	.321	3.026	.080	.311
	5	.564							
1000	2	.135	1	1.375	.033		1.261	.031	
	3	.220	2	.637	.042	.192	.555	.040	.174
	4	.625	3	1.874	.041	.274	2.020	.055	.265
	5	.342							
1500	2	.109	1	.824	.025		.766	.026	
	3	.189	2	.406	.037	.145	.307	.031	.142
	4	.442	3	1.093	.030	.225	1.338	.042	.217
	5	.269							
2000	2	.096	1	.455	.023		.485	.021	
	3	.161	2	.301	.029	.124	.255	.024	.123
	4	.363	3	.563	.028	.179	.901	.035	.185
	5	.212							
2500	2	.081	1	.406	.020		.408	.019	
	3	.137	2	.262	.026	.108	.217	.021	.096
	4	.331	3	.465	.024	.154	.651	.028	.163
	5	.188							
3000	2	.080	1	.354	.019		.353	.018	
	3	.127	2	.235	.023	.100	.196	.019	.092
	4	.309	3	.419	.022	.143	.526	.025	.143
	5	.182							
3500	2	.069	1	.319	.017		.324	.016	
	3	.110	2	.217	.021	.093	.189	.018	.085
	4	.263	3	.379	.020	.127	.390	.023	.128
	5	.164							
4000	2	.070	1	.297	.017		.309	.015	
	3	.112	2	.205	.020	.083	.171	.017	.077
	4	.267	3	.331	.019	.115	.370	.023	.125
	5	.163							
4500	2	.063	1	.280	.016		.280	.013	
	3	.110	2	.193	.019	.077	.159	.016	.073
	4	.241	3	.305	.018	.111	.339	.020	.113
	5	.150							
5000	2	.059	1	.262	.014		.267	.013	
	3	.100	2	.178	.019	.072	.151	.015	.071
	4	.242	3	.275	.017	.102	.322	.020	.111
	5	.147							

Note. [1] Median standard deviation of class-specific delta beta parameter estimates.

Table 3.11. Bias of standard error estimates (*bias*$_{SE}$)

	rmGPCM-3						mPCM-3		
N	**Item**	$\boldsymbol{\lambda^i}$	**Class**	$\boldsymbol{\Delta\beta_{0sg}}^1$	$\boldsymbol{\lambda_g}$	$\boldsymbol{\pi_g}$	$\boldsymbol{\Delta\beta_{0sg}}^1$	$\boldsymbol{\lambda_g}$	$\boldsymbol{\pi_g}$
500	2	.047	1	1.731	.006		1.863	.009	
	3	.066	2	1.072	.016	.086	.809	.021	.076
	4	.333	3	2.257	.013	.118	2.498	.031	.108
	5	.149							
1000	2	.015	1	.687	.003		.604	.004	
	3	.027	2	.235	.006	.051	.226	.010	.040
	4	.143	3	1.207	.007	.091	1.368	.017	.083
	5	.053							
1500	2	.011	1	.258	.001		.259	.003	
	3	.020	2	.093	.006	.024	.055	.005	.028
	4	.063	3	.560	.003	.063	.684	.010	.054
	5	.035							
2000	2	.008	1	.080	.002		.095	.002	
	3	.015	2	.042	.002	.016	.031	.002	.021
	4	.043	3	.173	.003	.032	.372	.007	.035
	5	.021							
2500	2	.005	1	.054	.001		.073	.001	
	3	.010	2	.027	.002	.010	.021	.002	.006
	4	.037	3	.110	.002	.021	.201	.003	.026
	5	.017							
3000	2	.007	1	.039	.001		.044	.002	
	3	.009	2	.022	.001	.010	.017	.001	.007
	4	.034	3	.073	.002	.020	.112	.003	.017
	5	.015							
3500	2	.004	1	.029	.001		.033	.001	
	3	.008	2	.019	.001	.008	.015	.001	.007
	4	.023	3	.056	.001	.013	.060	.002	.013
	5	.012							
4000	2	.006	1	.024	.002		.028	.001	
	3	.007	2	.016	.001	.005	.012	.001	.004
	4	.026	3	.037	.001	.008	.051	.003	.015
	5	.013							
4500	2	.004	1	.020	.001		.021	.000	
	3	.010	2	.015	.001	.004	.012	.001	.004
	4	.018	3	.033	.001	.010	.038	.002	.011
	5	.011							
5000	2	.003	1	.018	.000		.020	.001	
	3	.006	2	.012	.002	.003	.009	.001	.005
	4	.023	3	.027	.001	.007	.035	.002	.012
	5	.012							

Note. [1] Median *bias*$_{SE}$ for class-specific delta beta parameters.

Table 3.12. Width of the confidence interval for parameter estimates (*$width_{CI}$*)

	rmGPCM-3						mPCM-3		
N	**Item**	λ^i	**Class**	$\Delta\beta_{0sg}$[1]	λ_g	π_g	$\Delta\beta_{0sg}$[1]	λ_g	π_g
500	2	.712	1	3.062	.165		3.146	.147	
	3	1.249	2	2.049	.199	.664	1.817	.165	.638
	4	2.974	3	3.152	.185	.797	3.395	.195	.796
	5	1.792							
1000	2	.508	1	2.157	.119		2.199	.106	
	3	.876	2	1.448	.147	.556	1.308	.121	.525
	4	2.028	3	2.361	.134	.720	2.548	.149	.712
	5	1.249							
1500	2	.414	1	1.767	.097		1.778	.088	
	3	.700	2	1.211	.123	.479	1.073	.101	.451
	4	1.648	3	1.904	.112	.637	2.117	.127	.641
	5	.995							
2000	2	.361	1	1.511	.084		1.544	.077	
	3	.601	2	1.057	.108	.427	.935	.089	.402
	4	1.427	3	1.666	.098	.580	1.841	.110	.593
	5	.853							
2500	2	.320	1	1.342	.076		1.373	.069	
	3	.538	2	.943	.096	.388	.833	.078	.363
	4	1.257	3	1.494	.087	.526	1.659	.100	.539
	5	.759							
3000	2	.292	1	1.228	.069		1.248	.063	
	3	.493	2	.863	.088	.355	.760	.072	.335
	4	1.152	3	1.358	.080	.482	1.503	.091	.496
	5	.688							
3500	2	.266	1	1.141	.065		1.160	.059	
	3	.442	2	.801	.082	.337	.707	.068	.310
	4	1.017	3	1.247	.076	.452	1.356	.084	.454
	5	.636							
4000	2	.252	1	1.055	.060		1.083	.055	
	3	.425	2	.750	.076	.310	.658	.063	.290
	4	.995	3	1.171	.069	.420	1.292	.080	.431
	5	.597							
4500	2	.238	1	1.005	.057		1.022	.052	
	3	.396	2	.711	.072	.297	.621	.059	.274
	4	.905	3	1.090	.067	.399	1.191	.074	.402
	5	.564							
5000	2	.227	1	.949	.053		.965	.049	
	3	.381	2	.676	.067	.278	.587	.056	.262
	4	.881	3	1.037	.061	.376	1.148	.071	.388
	5	.539							

Note. [1] Median w*$idth_{CI}$* for class-specific delta beta parameters.

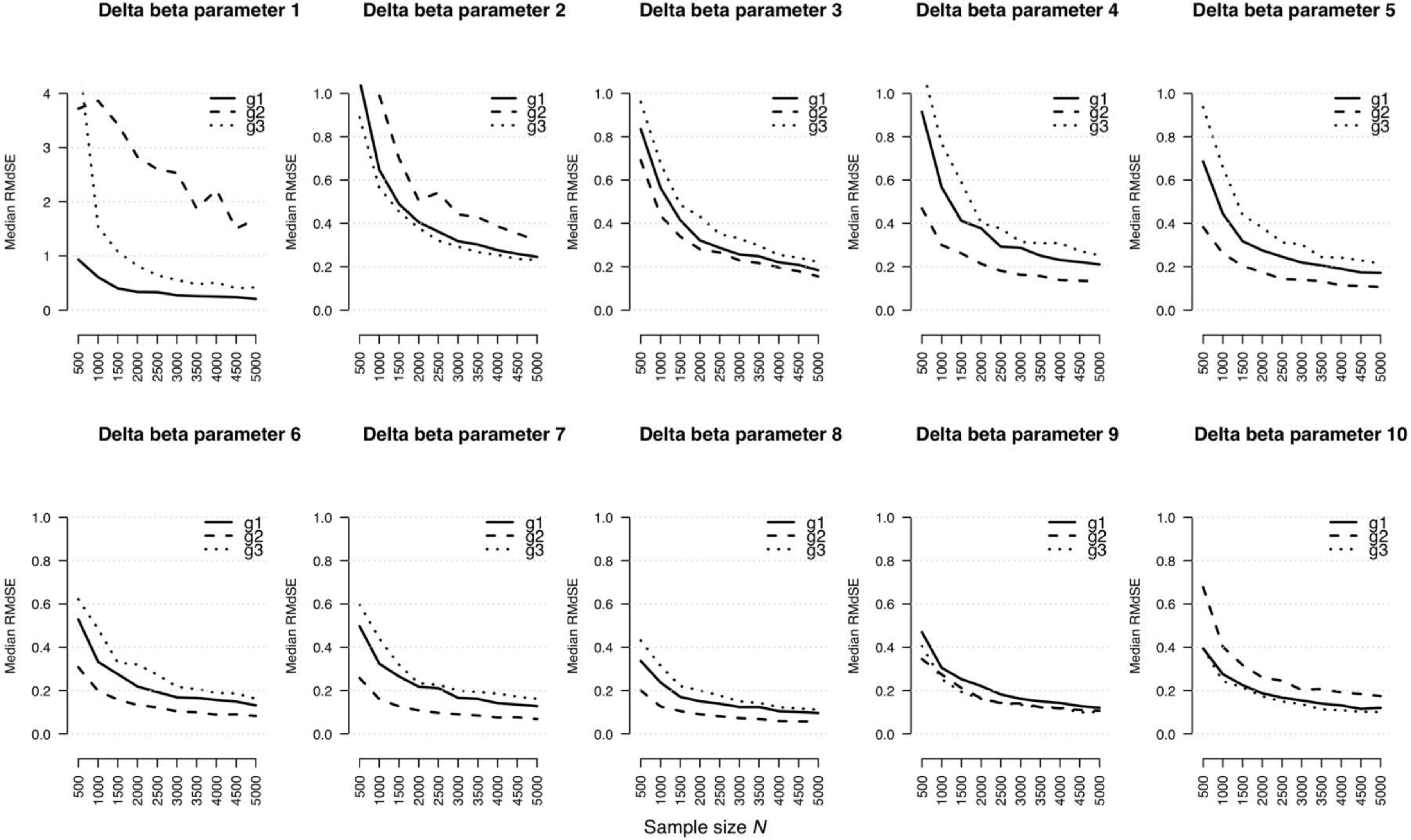
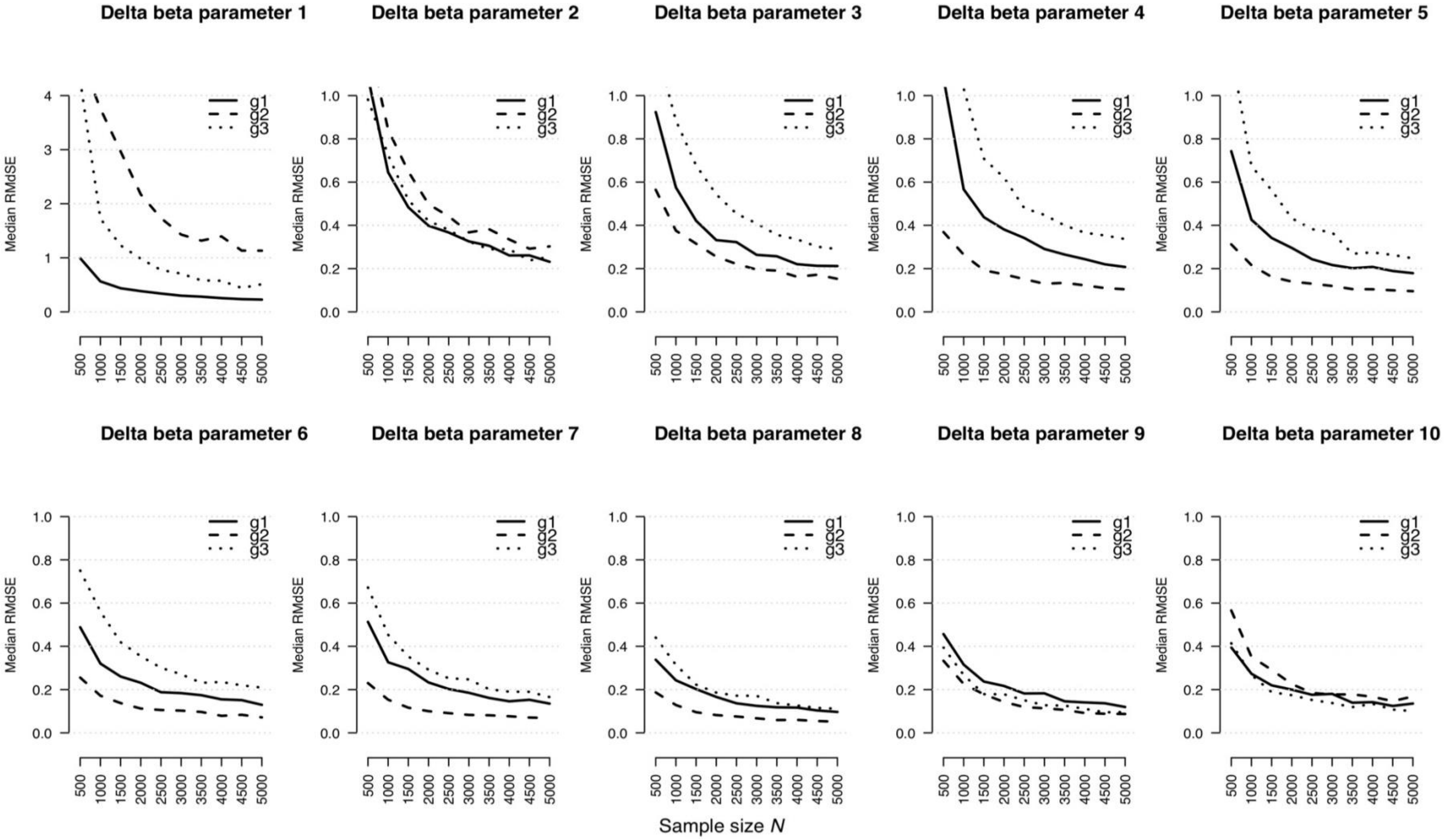


Figure 3.4. Root median squared error for class-specific delta beta parameter estimates in the rmGPCM-3.


Delta beta parameter 1

Delta beta parameter 2

Delta beta parameter 3

Delta beta parameter 4

Delta beta parameter 5

Delta beta parameter 6

Delta beta parameter 7

Delta beta parameter 8

Delta beta parameter 9

Delta beta parameter 10

Figure 3.5. Root median squared error for class-specific delta beta parameter estimates in the mPCM-3.

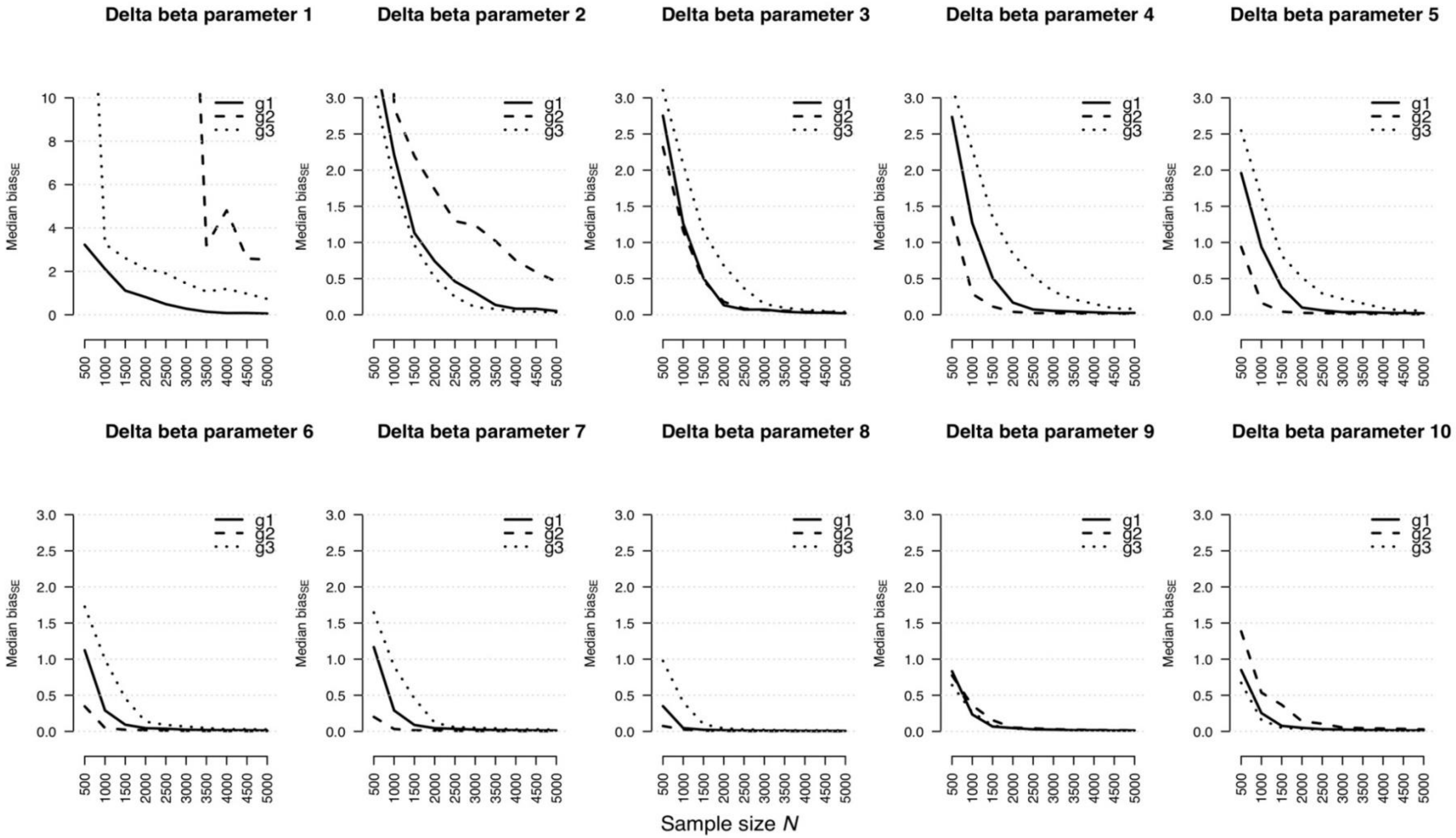


Figure 3.6. Bias of standard error estimates for class-specific delta beta parameter estimates in the rmGPCM-3.

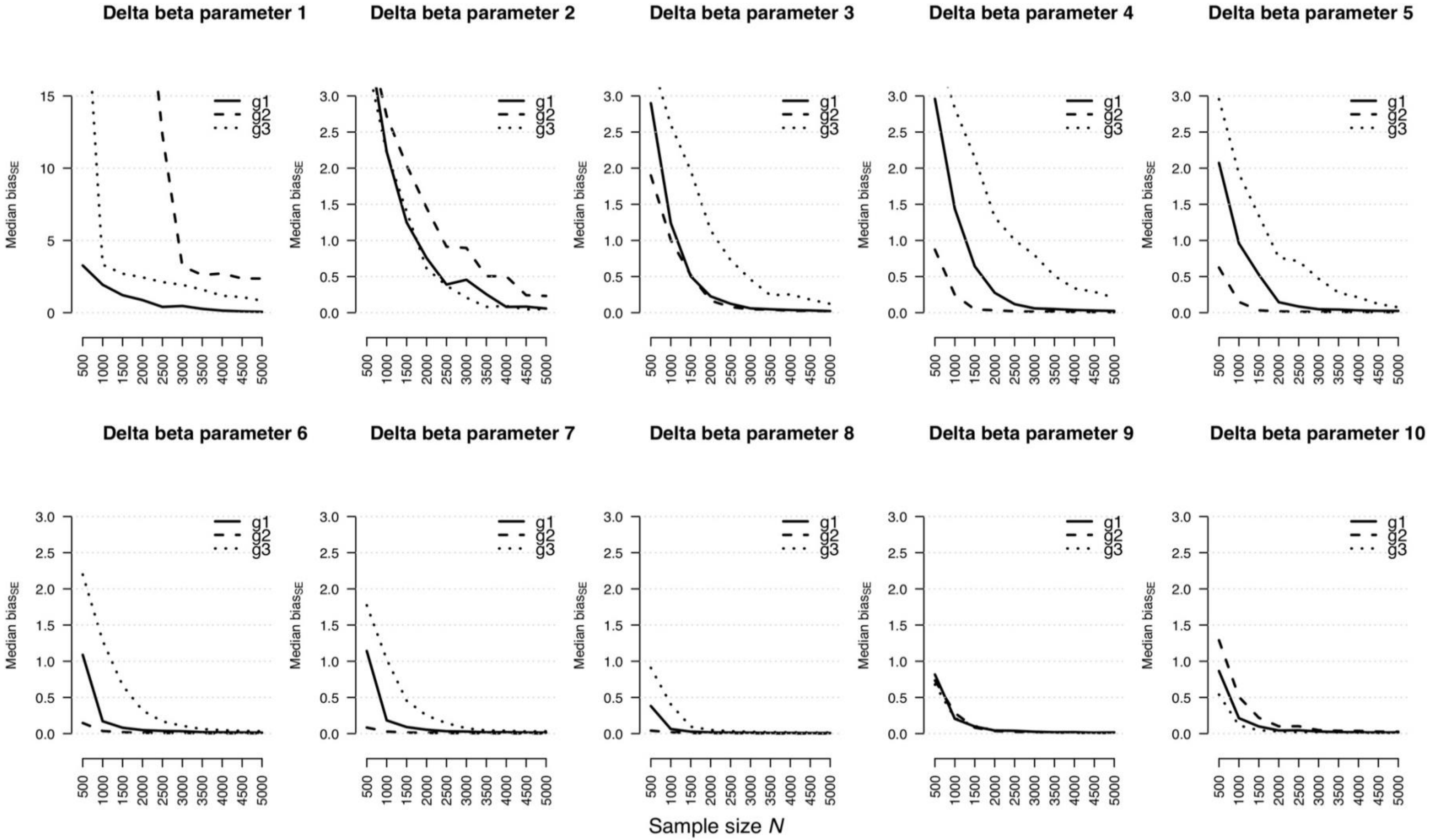


Figure 3.7. Bias of standard error estimates for class-specific delta beta parameter estimates in the mPCM-3.

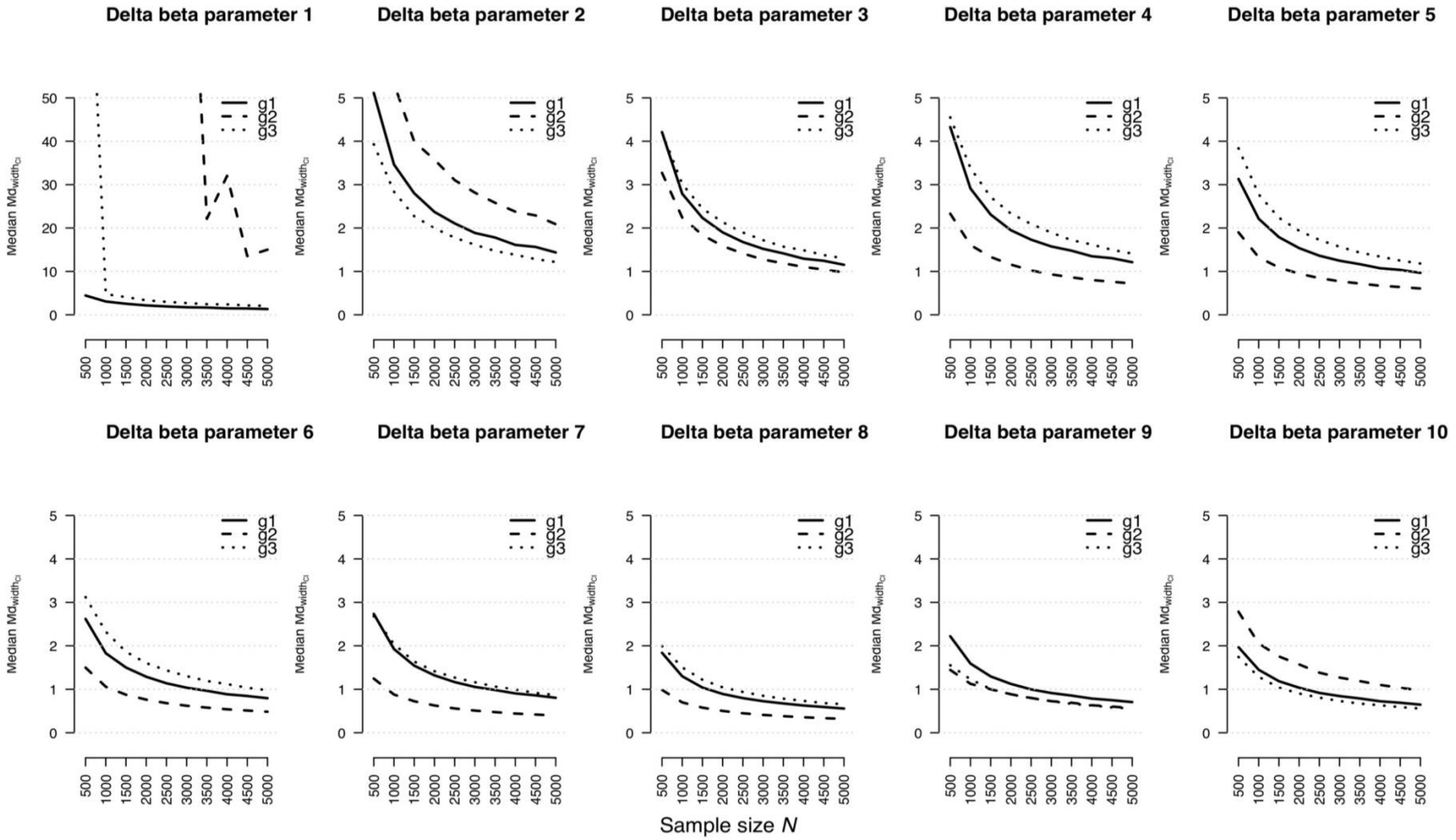


Figure 3.8. Width of the confidence interval for class-specific delta beta parameter estimates in the rmGPCM-3.

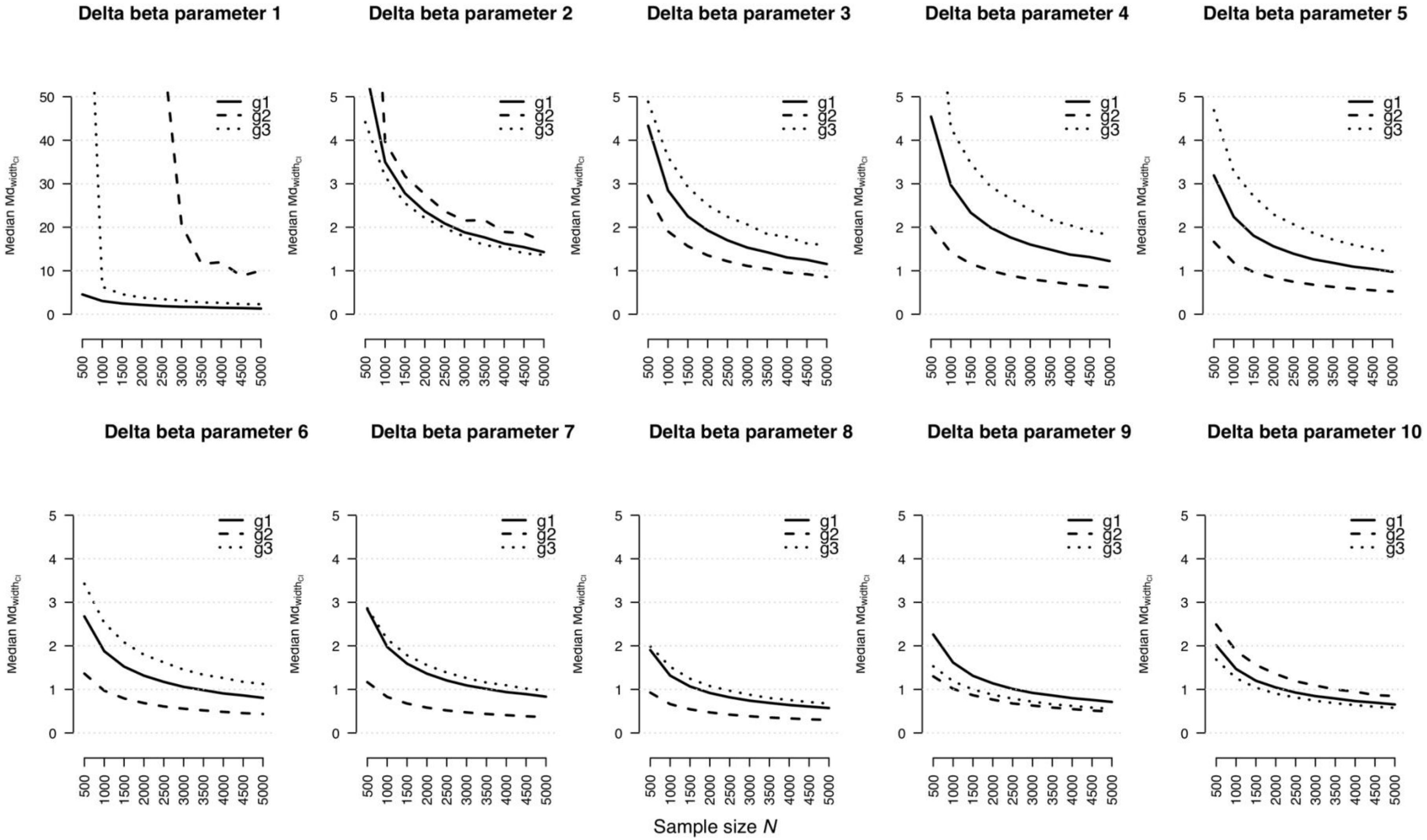


Figure 3.9. Width of the confidence interval for class-specific delta beta parameter estimates in the mPCM-3.

Table 3.13. Expected category probabilities in the latent classes for the population rmGPCM-3.

		Category										
Item	**Class**	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	1	.023	.009	.019	.032	.035	.116	.063	.139	.177	.051	.334
2	1	.021	.009	.019	.018	.018	.0645	.023	.040	.105	.053	.605
3	1	.002	.002	.005	.009	.01	.058	.042	.095	.176	.069	.519
4	1	.001	.001	.004	.007	.010	.073	.042	.101	.215	.051	.496
5	1	.004	.003	.007	.010	.013	.065	.021	.072	.140	.054	.612
1	2	.0004	.004	.019	.044	.066	.125	.160	.252	.278	.041	.013
2	2	.001	.005	.016	.037	.036	.081	.084	.152	.306	.156	.127
3	2	0	0	.003	.016	.029	.084	.123	.241	.380	.096	.029
4	2	0	.0001	.003	.015	.044	.125	.164	.295	.314	.034	.006
5	2	.0002	.002	.010	.028	.041	.095	.128	.230	.307	.114	.045
1	3	.004	.019	.034	.048	.038	.087	.088	.205	.211	.234	.034
2	3	.013	.017	.021	.014	.022	.060	.024	.115	.144	.376	.195
3	3	.001	.006	.010	.015	.015	.042	.058	.138	.170	.443	.103
4	3	0	.003	.006	.013	.016	.071	.088	.164	.186	.401	.053
5	3	.002	.007	.014	.018	.025	.071	.052	.097	.167	.365	.181
Mean	1	.010	.005	.011	.015	.017	.075	.038	.089	.162	.056	.521
Mean	2	.000	.002	.010	.028	.043	.102	.132	.234	.322	.088	.044
Mean	3	.004	.010	.017	.021	.023	.066	.062	.144	.175	.364	.113

Table 3.14. Expected category probabilities in the latent classes for the population mPCM-3.

							Category					
Item	**Class**	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	1	.012	.007	.016	.028	.033	.117	.065	.145	.184	.052	.337
2	1	.076	.004	.011	.012	.014	.057	.022	.042	.117	.058	.655
3	1	.002	.002	.005	.009	.010	.055	.042	.093	.177	.067	.538
4	1	.067	.003	.013	.016	.016	.090	.040	.085	.176	.044	.510
5	1	.010	.005	.010	.012	.015	.068	.019	.065	.127	.049	.620
1	2	.001	.004	.016	.041	.062	.129	.158	.255	.271	.051	.014
2	2	.0003	.002	.009	.025	.028	.074	.079	.158	.309	.178	.138
3	2	.0001	.0002	.004	.018	.031	.085	.120	.236	.359	.108	.039
4	2	0	.002	.013	.040	.069	.147	.152	.253	.285	.034	.005
5	2	.001	.005	.021	.045	.054	.107	.124	.214	.281	.108	.041
1	3	.002	.013	.028	.042	.036	.079	.094	.217	.226	.229	.035
2	3	.006	.009	.014	.010	.020	.058	.029	.124	.163	.35	.183
3	3	.001	.005	.009	.012	.011	.037	.054	.135	.180	.458	.099
4	3	.0002	.012	.014	.016	.011	.054	.065	.124	.173	.447	.083
5	3	.004	.009	.014	.013	.017	.060	.041	.077	.153	.391	.223
Mean	1	.008	.004	.011	.015	.018	.078	.038	.086	.156	.054	.532
Mean	2	.000	.003	.013	.034	.049	.109	.126	.223	.301	.096	.047
Mean	3	.002	.009	.016	.018	.019	.058	.057	.135	.179	.382	.124

4 Optimal Number of Response Categories for Assessing Job Satisfaction with a Rating Scale: an Experimental Online Study

Kutscher, T., & Eid, M. (2018). *Optimal Number of Response Categories for Assessing Job Satisfaction with a Rating Scale: an Experimental Online Study*. Manuscript in revision.

Abstract

When job satisfaction is measured in national panel surveys using a rating scale consisting 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 the 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 gather evidence concerning the optimal rating scale length. A sample of MTurk workers (*N* = 6,999) filled out a 12-item online questionnaire on aspects of job satisfaction, with a 4, 6, or 11 ordered response categories randomly assigned. The extent of response styles and reliability are used as evaluation criteria. 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. Considering the three-dimensional structure of the job satisfaction measure, we apply a multidimensional extension of the restrictive mixed generalized partial credit model to explore category use patterns within each condition. The results show similar configuration of three response-style classes in all conditions. Nevertheless, the proportion of respondents who use the rating scale inappropriately was lower in conditions with fewer response categories. An exception is the extreme response style, which showed a similar prevalence rate in all conditions. Furthermore, we find 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. Based on these results and model-based reliability scores, we conclude that a 6-point rating scale, followed by a 4-point rating scale, is the most adequate rating scale for assessing job satisfaction. 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

Optimal Number of Response Categories for Assessing Job Satisfaction with a Rating Scale: an Experimental Online Study

Job satisfaction (JS), as a component of subjective well-being, is a standard indicator of quality of life (Diener & 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 on 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 JS indicates areas of concern that require a manager's attention (Judge, Thoresen, Bono, & Patton, 2001; Tooksoon, 2011). For individuals, JS is one of the important areas of life that affects individual well-being and life satisfaction (Bowling, Eschleman, & Wang, 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, Bevitt, Freidin, Hahn, La, Macalalad, ... & Wooden, 2017), the German Socio-Economic Panel (GSOEP; Wagner, Frick, & Schupp, 2007), or the Swiss Household Panel Survey (SHP; Voorpostel, Tillmann, Lebert, Weaver, Kuhn, Lipps, ... & Wernli, 2010), JS is assessed using a rating scale consisting 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 response style [MRS], acquiescence [ARS], or discquiescence [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 & Steenkamp, 2001; Van Vaerenbergh & Thomas, 2012). The cause may be an increased response burden for respondents due to difficulties they experience in mapping their subjective categories in which they think about a certain construct to the rating scale if it includes an excessively large number of response categories (Baumgartner & Steenkamp, 2001; Cox, 1980; Viswanathan, Sudman, & Johnson, 2004). Should interindividual differences in RSs exist in the data, they will lower the data quality and can result in the validity of conclusions drawn from panel studies being called into question (Baumgartner & Steenkamp, 2001; De Jong, Steenkamp, Fox, & Baumgartner, 2008; Morren, Gelissen, & Vermunt, 2012; Podsakoff, MacKenzie, Lee, & Podsakoff, 2003). However, to date, no experimental study has compared the appropriateness of an 11-point rating scale with that of shorter rating scales with respect to a valid assessment of JS.

Using a between-subject experimental design, this paper aims to identify an adequate rating scale for assessing aspects of JS. This design makes it possible to detect whether rating scales consisting 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 include positively worded items written in plain language. In the following section, we summarize empirical evidence calling into question the adequacy of long rating scales with regard to collecting high-quality data. Thereafter, we present a literature review on the effects of rating scale length on RS. Due to a lack of consistent knowledge concerning the other major factor affecting RS, namely a stable individual response style, an influential factor that is independent of the features of a rating scale (Austin, Deary, & Egan, 2006; Billiet & Davidon, 2008; Kieruj & Moors, 2013; Krosnick, 1999; Van Vaerenbergh & Thomas, 2013), we also address the relevance of respondent characteristics to 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 and provide practical recommendations for the selection of an optimal rating scale.

4.1 Evidence against the Appropriateness of Long Rating Scales

First, the poor performance of long rating scales has been demonstrated in a series of empirical studies. By analyzing the JS data measured with an 11-point rating scale in the HILDA survey, Kutscher, Crayen, and Eid (2017) applied mixture models of the item response theory (IRT) and detected at least two serious shortcomings of this rating scale: (i) a very high proportion of respondents used 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 one-third of their respondents had a strong preference for specific response categories and tended to ignore only one category, and the majority of respondents (71%) used this rating scale in an appropriate manner. These findings illustrate the superior performance of short rating scales, which suggests that RSs may be partly avoided by shortening the rating scale used. 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, Tietjens, & Strauss, 2013; Khadka, Gothwal, McAlinden, Lamoureux, & Pesudovs, 2012).

Second, according to the response process model (see Tourangeau, Rips, & Rasinski, 2000), respondents report their judgments by selecting the matching category of the response format offered. This is one of the cognitive steps that they take when answering items. In this step, if the rating scale used is not adapted to the respondents' thinking complexity and ability to discriminate, it may prove to be a source of RSs (Baumgartner & Steenkamp, 2001; Cox, 1980; Viswanathan et al., 2004). According to Krosnick's concept of satisficing (1991), inappropriate responding (satisficing) of the part of 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 & Levine, 2016; Krosnick, 1999; Swait & Adamowicz, 2001; Viswanathan et al., 2004; Weathers, Sharma, & Niedrich, 2005). In this case, the poor performance of a long rating scale 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, Nash, & Gillmor, 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, Beal, & Payne, 2009; Weathers et al., 2005; Weijters, Cabooter, & Schillewaert, 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 (Weather 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, Baldueza, Barner-Rasmussen, Barzantny, Canabal, Davila, ... & Liang, 2009; Hui & Triandis, 1989). To conclude, from a cognitive point of view, excessively long rating scales (as well as excessively short rating scales) are clearly sub-optimal.

Third, long rating scales are often used in the social research due to the finding established in previous research that improved reliability can be attained by increasing rating scale length (Alwin & Krosnick, 1991; Preston & Colman, 2000; see for meta-analysis Churchill & Peter, 1984; Saris & 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 & Wang, 2014; Revilla, Saris, & Krosnick, 2014; Tarka, 2016). When the effects of RSs are controlled, reliability and criterion validity scores increase as a rating scale expands up to four or five response options; then 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 & Paek, 2014; Lozano, Carcìa-Cueto, Muñiz, 2008; Maydeu-Olivares, Kramp, García-Forero, Gallardo-Pujol, & Coffman, 2009). 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, the optimal rating scale length seems to range between four and six or seven categories.

4.2 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, Weijters, Geuens, & Vermeir, 2016; Hamby & Levine, 2016; Kieruj & Moors, 2010; Moors, 2008; Moors, Kieruj, & Vermunt, 2014; O'Muircheartaigh, Krosnick, & Helic, 1999; Tourangeau, Couper, & Conrad, 2007; Weijters et al., 2010). Regarding rating scale length, Harzing and colleagues (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 & 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) are sub-optimal due to 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 confirms 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 & Stachowski, 2013; 2009). 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, Drasgow, & González-Romá, 2004; Presser & Schuman, 1980; Sturgis, Roberts, & Smith, 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, Hox, & Sikkel, 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 due to a deficient response format.

4.3 Respondent Characteristics for Predicting 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 (A Austin et al., 2006; Gerber-Braun, 2010; Kieruj & Moors, 2013; Meiser & Machunsky, 2008), low in conscientiousness (Zettler, Lang, Hülsheger, & Hilbig, 2015), high in neuroticism (Baumgartner & Steenkamp, 2001; Hernández et al, 2004), and low in openness to experience (Meiser & 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 & Machunsky, 2008). Moreover, for ERS use, research has highlighted the relevance of intolerance to 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 & Rauber, 2000; Weijters, Geuens, & Schillewaert, 2010b), whereas other studies have found education to have no effect (Kieruj & 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 & 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 & 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 & 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 only rudimentarily examined compared to the ERS, thus creating an inconsistent picture concerning individuals' use of RSs.

4.4 Research Questions and Expectations

The present research aims to identify an optimal rating scale for assessing JS, which is considered an important indicator of quality of life, by conducting a between-subject experiment with a varied 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 excluded 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 focused on two criteria: the extent of RSs present in the data and 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, applying 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 RSs and ignored response categories. 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 being found at different rating scale lengths. Therefore, in contrast to previous research, which has mostly focused on a few selected predictors, this research 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. In particular, this paper provides recommendations concerning the optimal number of response categories for assessing JS using a rating scale.

4.5 Methods

4.5.1 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, Kwang, & Gosling, 2011; Mason & 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, having an employment relationship, and living in the USA. Additionally, to reduce the risk of satisficing response behavior, we followed the recommendations by Peer, Vosgerau, and Acquisti (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

minutes (*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 6,999 MTurk workers, all of whom filled out the online questionnaire without any gaps. According to current simulation findings (e.g., Cho, 2013; Huang, 2016; Jin & Wang, 2014; Kutscher, Crayen, & Eid, submitted), this sample size (namely, 2,000–2,500 individuals per condition) should be sufficient for the mixed polytomous IRT model to display optimal performance. We applied this model within each of experimental conditions to explore category use (see the section titled "Multidimensional restricted mixed IRT model" below). 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). 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%).

4.5.2 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 4.1). The experimental groups did not differ in terms of age (*F*(2, 6995) = 0.05, *p* = .951), gender (χ^2(2) = 4.35, *p* = .114), education (χ^2(6) = 4.90, *p* = .557), employment type (χ^2(4) = 4.87, *p* = .301), or leadership level (χ^2(8) = 6.95, *p* = .542). These results were obtained by means of a univariate analysis of variance (ANOVA) for the continuous variable age and a Pearson's chi-square test for the nominal variables.

Table 4.1. Comparison of sample characteristics among experimental conditions.

	11 categories (*N* = 2,322)	**6 categories (*N* = 2,364)**	**4 categories (*N* = 2,313)**
Age (Mean, SD)	34.08 (10.95)	34.05 (11.13)	33.98 (11.19)
Gender (female) (%)	61.74	61.95	59.26
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, MTurk worker	24	23	24
Part-time occupation (%)	26	29	28

4.5.3 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 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 study and the HILDA surveys). Respondents answered the JS items with an

11-, 6-, or 4-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 & 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 & John, 2007). The dimensions consisted 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. The reliabilities of the subscales were acceptable with regard to short subscale length (McDonald's ω = .68, .66, .50, .56, and .52 for extraversion, neuroticism, openness to experience, conscientiousness, and agreeableness subscales, respectively).

The General Self-Efficacy Scale

The general self-efficacy scale (GSE; Schwarzer & 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*). The reliability (McDonald's ω) was .89.

Tolerance to Ambiguity

Tolerance to ambiguity (TA) was assessed with a set of six items selected from the original Ambiguity Tolerance Scale (AT-20; MacDonald, 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*). Similar to the original scale, the short TA scale showed a unidimensional latent structure: the one-factor confirmatory factor analysis (CFA) model demonstrated an acceptable model fit (χ^2(9) = 366.97, p < .001; RMSEA = .08, 90%-CI [0.07; 0.08]; CFI = .94; TLI = 0.90; SRMR = 0.05). See Table 4.7 in the appendix to Chapter 4. The reliability was acceptable (McDonald's ω = .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, Petty, Feinstein, & Jarvis, 1996) based on whether they had item 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 (χ^2 (35) = 862.07, *p* < .001; RMSEA = .06, 90%-CI [0.055; 0.062]; CFI = .98; TLI = 0.98; SRMR = 0.05; for details, see Table 4.8 in the appendix to Chapter 4). 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 ω for the reduced scale was .89.

Decisiveness scale

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*). The reliability score (McDonald's ω) was .83.

Social Desirability

Social Desirability was assessed using a short version of the Balanced Inventory of Desirable Responding (Winkler, Kroh, & Spiess, 2006), which measures the 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 turn 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 reliabilities of the two subscales were acceptable when taking into consideration the short length of these subscales (McDonald's ω = .80 and .61 for the SDE and the IM, respectively).

Verbal memory ability

Verbal memory ability was assessed using 10 questions randomly selected from the original 20-question

Verbal Memory Measure which is a part of the Intelligence Structural Test (the English version of IST 2000 R; Beauducel, 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, respondents were posed questions (e.g., "The word with the initial letter – M – was a/an ...?") that prompted them to select only one 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. 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; Beauducel, 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. 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 (*at 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 and colleagues (2017), was verified in this study by carrying out the CFA, which indicated an acceptable appropriate model fit (χ^2 (29) = 1665.85, p < .001; RMSEA = .09, 90%-CI [0.09; 0.09]; CFI = .93; TLI = 0.90; SRMR = 0.07; for details, see in Table 4.9 in the appendix to Chapter 4). 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* (level 1: manager, self-employed, etc.; level 2: professional, researcher, etc., level 3: technician, marketing, personal service worker, etc.; level 4: administrative worker, etc.; 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 (small: less than 50 employees, middle: 50–200 employees, large: 200 employees or more).

Sociodemographic 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").

4.6 Statistical Analyses

4.6.1 Multidimensional Restricted Mixed IRT Model

To analyze individual differences in RSs with respect to the JS items, the mixture IRT approach, which assumes the existence of latent classes of respondents with homogeneous response patterns, was applied within experimental conditions (for an overview, see von Davier & Yamamoto, 2007). 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 e.g., Weijters, Geuens, & Schillewaert, 2010a; Wetzel, Carstensen, & Böhnke, 2013; Zettler et al., 2015), it was assumed that the number of latent classes with a specific category use pattern is the same across latent dimensions. 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:

$$\mathrm{P_{vtix}}(\theta) = \sum_{g=1}^{G} \pi_g \frac{\exp\,[\sum_{s=0}^{x} \delta_{it}(\theta_{vtg} - \tau_{istg})]}{\sum_{c=0}^{m} \exp\,[\sum_{s=0}^{c} \delta_{it}(\theta_{vtg} - \tau_{istg})]}, \tag{4.1}$$

where $\mathrm{P_{vtix}}(\theta)$ denotes the probability of obtaining a response in a category *x* (*x* ∈{0,…, *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 θ_{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*. δ_{it} is a class-invariant discrimination parameter of item *i* belonging to dimension *t* (with $\delta_{it} > 0$ and with $\delta_{1t} = 1$), and τ_{istg} is a class-specific threshold parameter of item *i* with respect to dimension *t* (with *s* ∈{0,…, *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 (π_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 & 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 the large sample size (Dziak, Coffman, Lanza, & Li, 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. submitted). 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 χ^2 goodness-of-fit statistics or the likelihood-ratio test). The reason for this was the disproportionally high number of potential response patterns (e.g., 161,051 possible response patterns with 11 response categories for each of the five items) in relation to the sample size used in each of experimental conditions (range: 2,313–2,364 individuals). In addition, bootstrapping the distribution of the Pearson and Cressie-Read χ^2 goodness-of-fit statistics was unrealistic due to excessively lengthy computing times caused by the complexity of the model (Nylund, Asparouhov, & Muthén, 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 & Joe, 2008).

4.6.2 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, …, 5) in Figure 4.1, 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 (Figure 4.1), 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, Ying, & Brown, 2011; Wetzel & 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 4.1 represents the case of an item with nearly equidistant categories.

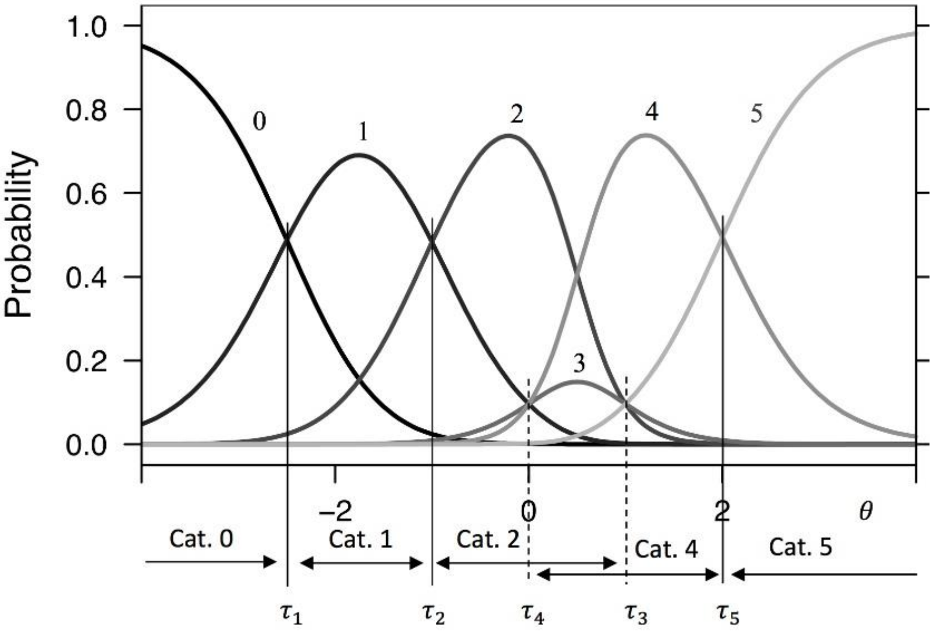


Figure 4.1. Category characteristic curves (CCCs) for a fictitious item with six response categories and an unordered threshold. (Threshold parameters: τ_1 = - 2.5, τ_2 = -1, τ_3 = 1, τ_4 = 0, τ_5 = 2; discrimination parameter δ = 2.)

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.

4.6.3 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., "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.

4.6.4 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 the applying the multidimensional rmGPCM to regression coefficients and standard errors (Vermunt & Magidson, 2013).

4.7 Results

4.7.1 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 & Moustaki, 2001) using the R package lavaan (Rosseel, 2012). This analysis revealed that the JS items had an oblique three-dimensional structure in all experimental conditions (see Table 4.10 in the appendix to Chapter 4). Table 4.2 presents the JS subscales and descriptive statistics for the JS items under different experimental conditions. Independent of rating scale length, 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 the 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 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 Figure 4.41, which can be found in the appendix to Chapter 4). Under the experimental conditions, the values of reliability (McDonald's ω) for all subscales were acceptable, but they decreased as the rating scale was shortened.

4.7.2 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 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 4.3). 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 show any clearly identifiable category use patterns but did include 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 also under this condition.

In addition, Table 4.3 indicates that the mean assignment probabilities and the estimated reliabilities are sufficiently large for all conditions. The values of both statistics were slightly smaller in conditions with shorter response formats.

Table 4.2. Descriptive statistics for the items of job satisfaction and the reliability scores of subscales under the different experimental conditions.

	11 categories			**6 categories**			**4 categories**		
Item	**Mean (SD)**	**Skew.**	**Kurt.**	**Mean (SD)**	**Skew.**	**Kurt.**	**Mean (SD)**	**Skew.**	**Kurt.**
Satisfaction with benefits and prospects		ω =.80			ω =.78			ω =.74	
Item 1. Your total pay	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. The 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. Your 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. The 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
Satisfaction with work tasks and conditions		ω =.80			ω =.76			ω =.73	
Item 5. The work tasks themselves	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. The hours you work	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. The 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. The flexibility to balance work and non-work commitments	6.94 (2.79)	-0.80	-0.23	3.53 (1.42)	-0.82	-0.17	2.09 (0.93)	-0.69	-0.52
Satisfaction with social aspects		ω =.87			ω =.85			ω =.81	
Item 9: The relation and communication with your immediate 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, recognition, and rewards for your 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. The work atmosphere and relations with your co-workers	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. Internal staff rules and regulations in your 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. Skew. = skewness; Kurt. = kurtosis; ω = McDonald's omega. The numerical value of the lowest category is always zero.

Table 4.3. Goodness-of-fit statistics for the multidimensional rmGPCM in experimental conditions.

Condition	**Model**	n_{par}	**LL**	**BIC**	**Mean assignment probability in classes**[1]	**Model-based reliability estimates for JS subscales**[2]
11 categories	1 class	135	-56,025	113,097		
	2 classes	259	-54,080	110,168		
	3 classes	383	-53,486	**109,941**	0.91, 0.92, 0.89	0.84, 0.84, 0.86
	4 classes	507	-53,197	110,323		
	5 classes	631	-52,953	110,796		
6 categories	1 class	75	-40,368	81,318		
	2 classes	139	-39,214	79,508		
	3 classes	203	-38,937	**79,450**	0.88, 0.87, 0.85	0.80, 0.79, 0.83
	4 classes	267	-38,768	79,611		
	5 classes	331	-38,640	79,852		
4 categories	1 classes	51	-30,045	60,485		
	2 classes	91	-29,349	**59,403**	0.88, 0.87	0.76, 0.74. 0.77
	3 classes	131	-29,222	59,458	0.87, 0.80, 0.80	0.75, 0.73, 0.77
	4 classes	171	-29,122	59,569		
	5 classes	211	-29,046	59,726		

Notes. [1] The values are reported in the following order: for the large class, middle-sized class, and small class.

[2] 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".

4.7.3 Category Use Patterns in Varied Number of Response Categories

Table 4.4 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 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 ignore 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, at the same time, 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 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 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 condition"), 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.

4.7.3.1 Class-Specific Category Use for the "Satisfaction with Work Tasks and Conditions" Subscale

Figure 4.2 presents the class-specific CCCs for the three different rating scales. (For the estimated item parameters used to draw these CCCs, see Table 4.11–4.13 in the appendix to Chapter 4.) For the 11-category condition (Figure 4.2a), 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.

Table 4.4. Number of unordered thresholds, category widths, and scale range for the rating scale including 11, 6, and 4 response categories.

	11 categories[1]			**6 categories[1]**			**4 categories[1]**		
	Benefits & prosp.	**Work and conditions**	**Social aspects**	**Benefits & prosp.**	**Work and conditions**	**Social aspects**	**Benefits & prosp.**	**Work and conditions**	**Social aspects**
Number of unordered thresholds per item									
Class 1[2]	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[2]	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[2]	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 brackets (in logits)									
Class 1[2]	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. [1] An item with 11, 6, and 4 response categories has 10, 5, and 3 thresholds, respectively.

[2] Latent classes within experimental conditions are sorted in descending order by their size: Class 1 is the largest, 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.

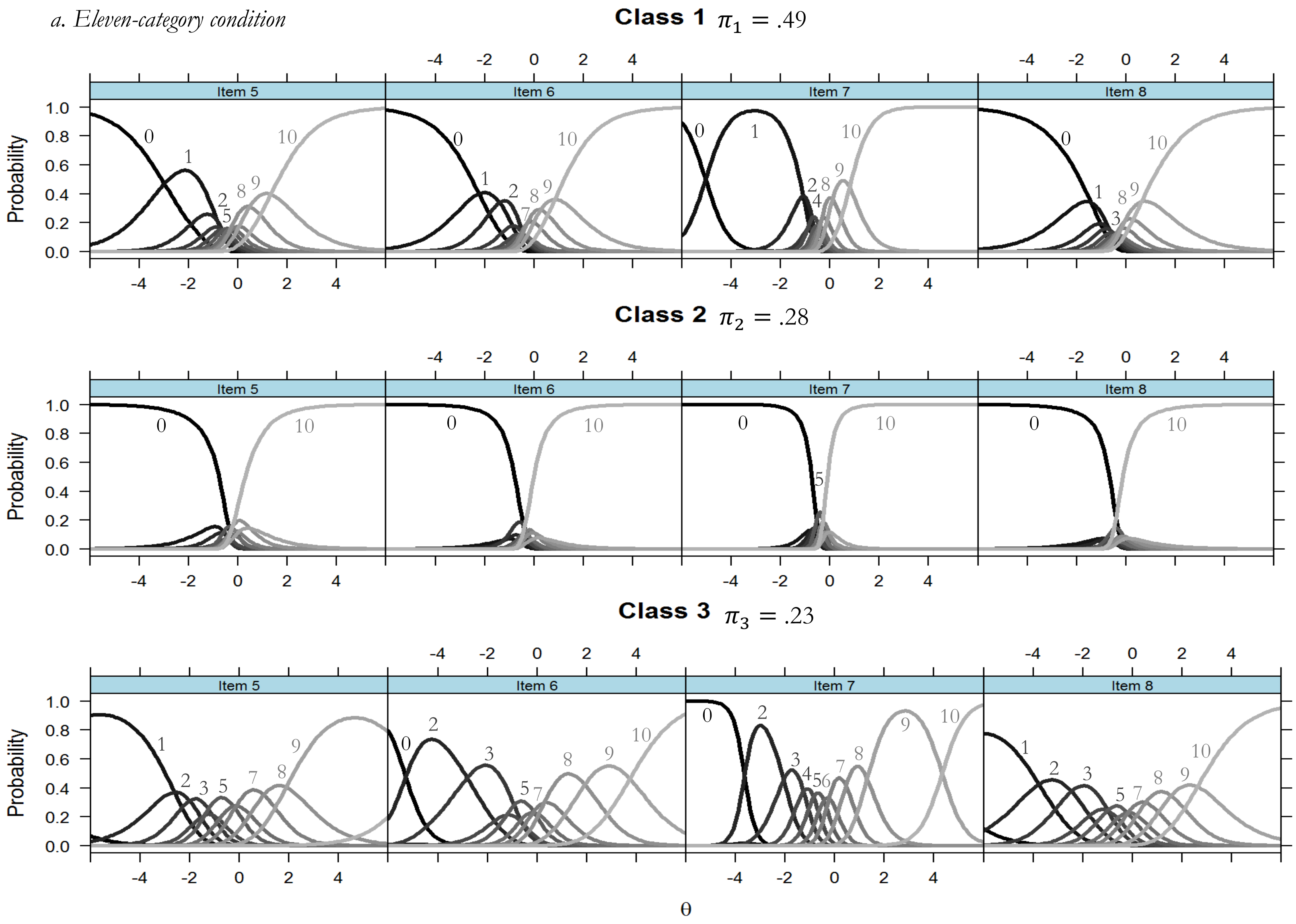
The CCCs of the middle categories ($x = 4$ to $x = 7$) are close to each other, indicating that these categories are avoided or that only one of them covers a very small segment of the latent continuum. The other half of the sample consisted of two latent classes. The medium-sized second class (28%) was characterized by using only the lowest and the highest response categories; all the categories 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 $\theta_{\text{vtg}} = -4$ and above $\theta_{\text{vtg}} = 4$). This means that the latent trait values of both very dissatisfied and very satisfied respondents were unreasonably low or large. 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 (Figure 4.2b), 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, the middle category $x = 3$ of which was ignored). However, compared to other 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, the middle category ($x = 3$). 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 (Figure 4.2c), 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.

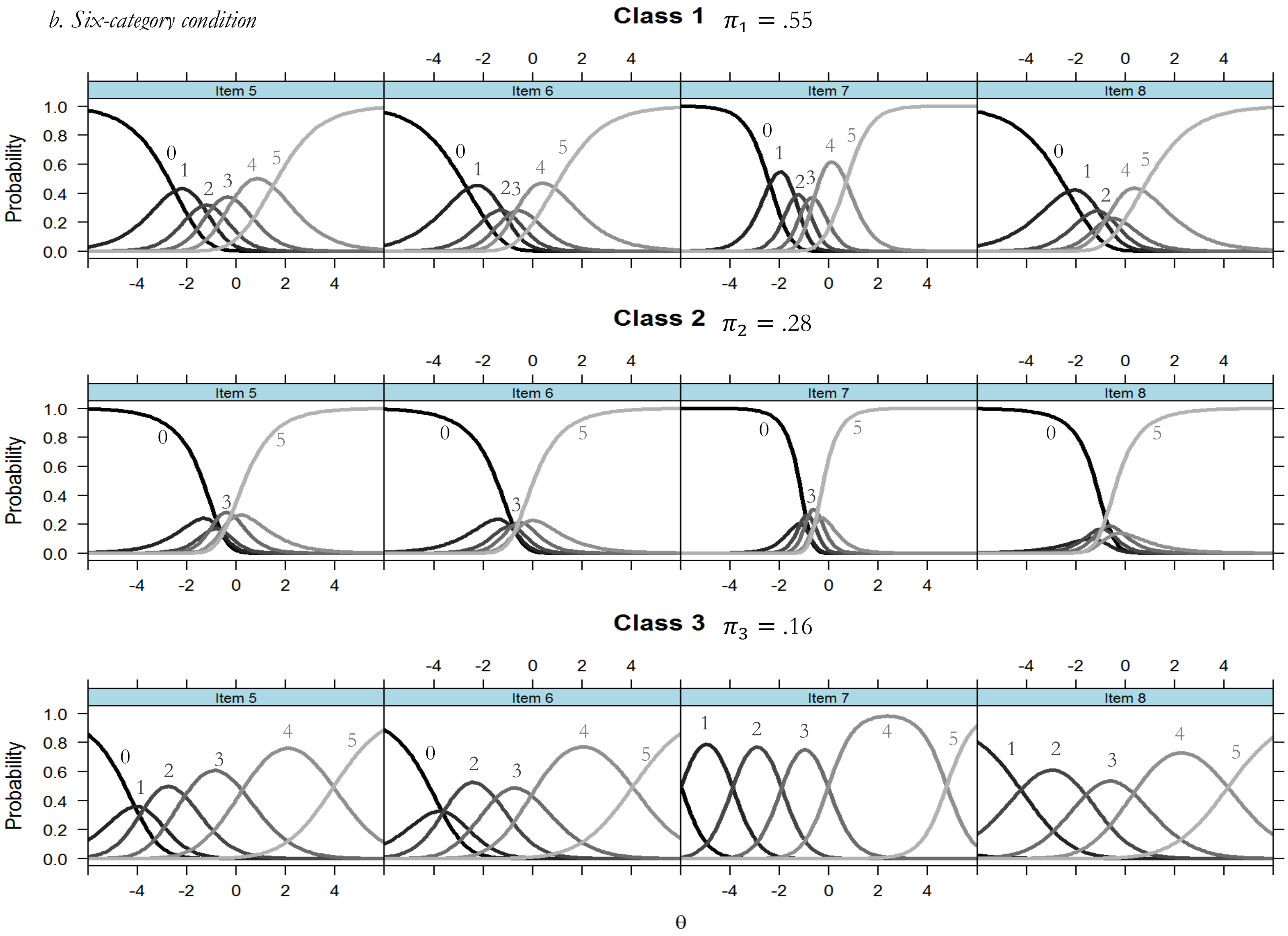
4.7.3.2 Specific Features of the Other Job Satisfaction Subscales

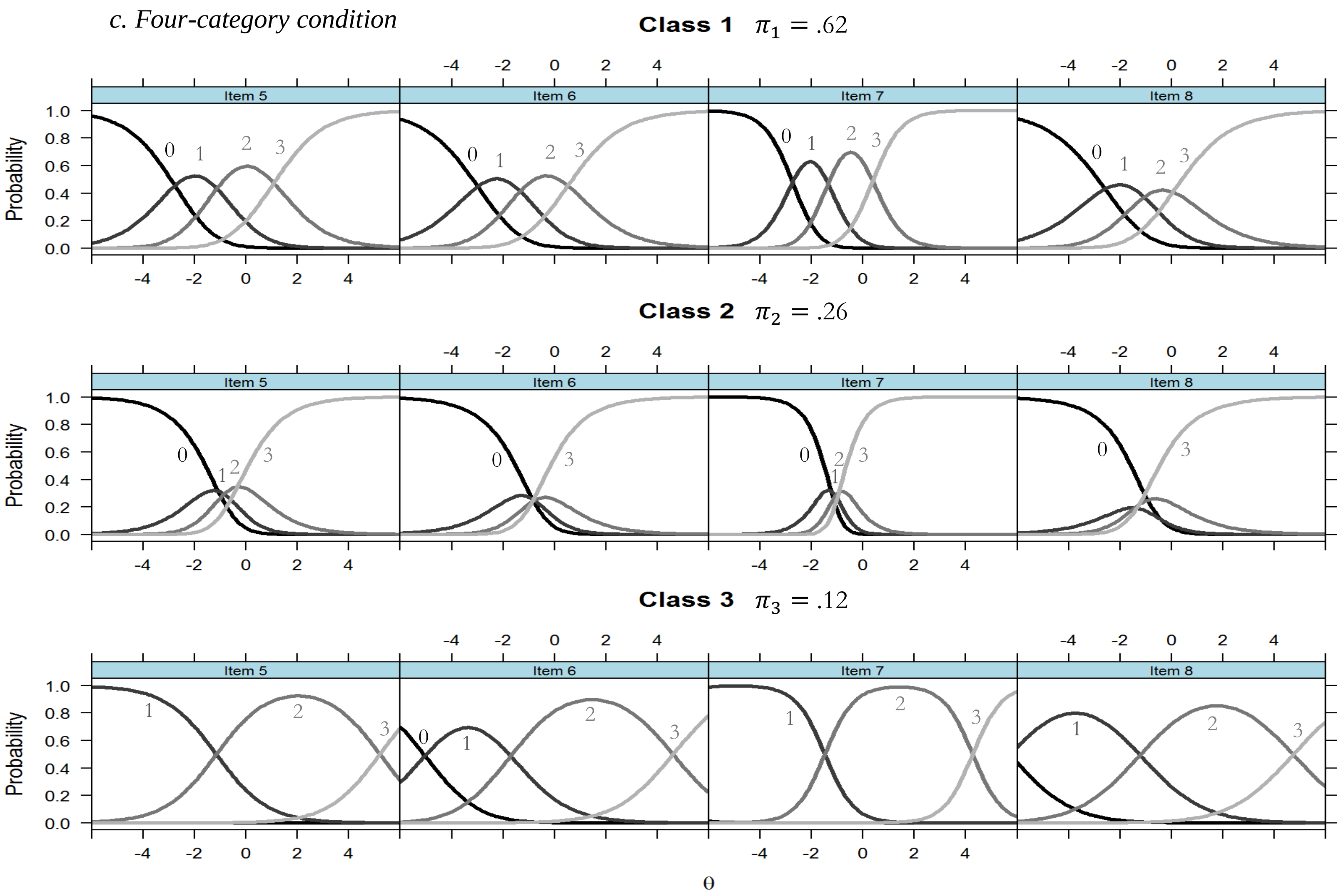
By and large, the patterns of class-specific category use within the various experimental conditions were similar across the JS subscales (see Figure 4.11–4.13 in the appendix to Chapter 4). However, a few specific features could be identified for the subscale "Satisfaction with benefits and prospects." For example, the ORS class ignored the middle categories ($x = 4$ to $x = 6$) in the 11-category condition



a. Eleven-category condition

b. Six-category condition





c. Four-category condition

Figure 4.2. 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.)

and for items 3 and 4 in the 6-category condition. In the ERS class, a middle category (if presented) covered a slightly larger segment of the latent continuum, as was observed for the subscale "Satisfaction with work tasks and conditions." 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 the middle categories. In the ORS class, we found that the middle category covered a larger segment of the latent trait variable in the 6-category condition. In the ERS class, the middle category had the highest probability of being endorsed only on a small area (with the exception of the 4-category condition).

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 scale length. However, the proportion of respondents who exhibited ORS increased with a reduction of the number of response categories (for the 11-category condition, 49%; for the 6-category condition, 55%; and, for the 4-category condition, 62%). Coincidently, 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 response format 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.

4.7.4 Expected Relative Category Frequencies of Job Satisfaction Items in Latent Classes

Figure 4.3 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 length (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 category 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

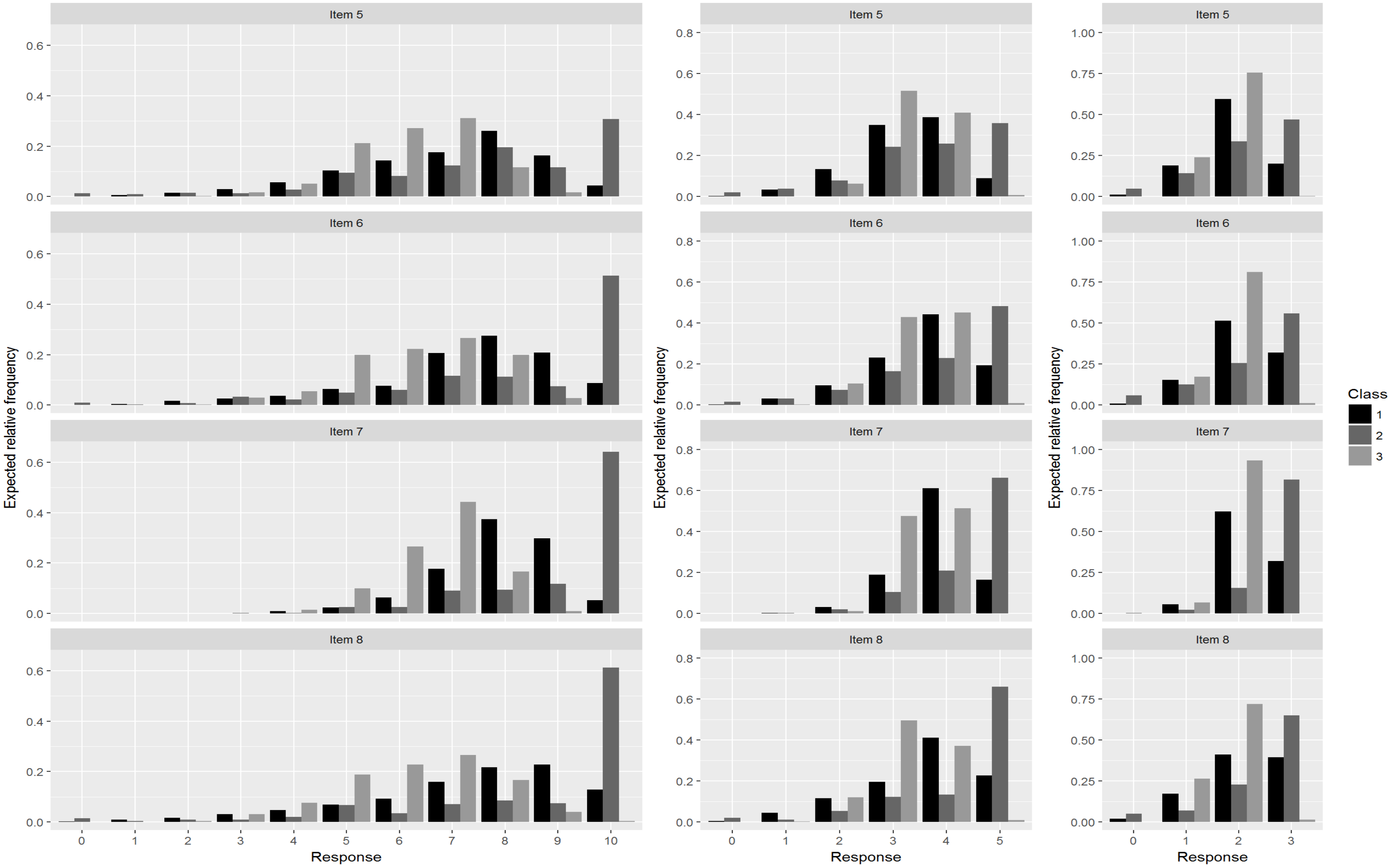


Figure 4.3. Expected relative frequencies for the items of the subscale "Satisfaction with work tasks and conditions" in the three experimental conditions.

(11-category condition in the left column, the 6-category condition in the middle column, and the 4-category condition in the right column.)

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 (Figure 4.7–4.8 in the appendix to Chapter 4).

4.7.5 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%); quick responding occurred less frequently (max. 4%; see Table 4.5). 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 hardly differed in the distribution of careless responding. However, within 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.

4.7.6 Prediction of Assignment to Latent Classes

Table 4.6 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 task and conditions and satisfaction with the 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 4.14 in the appendix to Chapter 4). For reasons of comprehensibility, Table 4.6

Table 4.5. Proportion of careless responding in latent classes within experimental conditions. (%)

Index	**11 categories** **(N = 2,322)**	**6 categories** **(N = 2,364)**	**4 categories** **(N = 2,313)**	**Test statistics**
Inattentive responding	6.8	6.2	6.2	Between conditions: $\chi^2(2) = 0.76, p = .68$
				Between classes within:
ORS class	5.7	5.4	5.1	4-category cond.: $\chi^2(2) = 28.59, p < .001$
ERS class	5.3	4.5	5.6	6-category cond.: $\chi^2(2) = 25.47, p < .001$
Non-ERS class	10.7	11.7	13.4	11-category cond.: $\chi^2(2) = 17.65, p < .001$
Quick responding	3.3	2.8	3.9	Between conditions: $\chi^2(2) = 4.43, p = .11$
				Between classes within:
ORS class	2.4	1.8	3.1	4-category cond.: $\chi^2(2) = 38.77, p < .001$
ERS class	1.5	3.3	2.6	6-category cond.: $\chi^2(2) = 13.81, p < .01$
Non-ERS class	7.2	5.2	10.6	11-category cond.: $\chi^2(2) = 35.42, p < .001$
Invariant responding	5.7	6.3	5.2	Between conditions: $\chi^2(2) = 2.49, p = .29$
				Between classes within:
ORS class	3.6	4.6	3.2	4-category cond.: $\chi^2(2) = 39.13, p < .001$
ERS class	6.5	6.0	7.4	6-category cond.: $\chi^2(2) = 28.08, p < .001$
Non-ERS class	9.2	11.9	11.3	11-category cond.: $\chi^2(2) = 22.44, p < .001$
Random responding	7.1	6.5	6.8	Between conditions: $\chi^2(2) = 0.65, p = .72$
				Between classes within:
ORS class	7.8	5.4	7.3	4-category cond.: $\chi^2(2) = 6.71, p < .05$
ERS class	9.1	9.8	7.2	6-category cond.: $\chi^2(2) = 16.03, p < .001$
Non-ERS class	3.0	4.7	3.2	11-category cond.: $\chi^2(2) = 19.12, p < .001$

presents only significant predictors and test statistics.[10] Independent of the response format, 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 a 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 4.6, these predictors are marked with a gray background.

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 4.6). 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 were open to new experiences. It was less likely for respondents with invariant

[10] All results can be obtained from the first author.

Table 4.6. Prediction of assignment to latent classes by means of multinomial regression models in the three experimental conditions.

	11 categories				6 categories				4 categories			
	B	**(SE)**	***e^b***	**95%-CI**	***B***	**(SE)**	***e^b***	**95%-CI**	***B***	**(SE)**	***e^b***	**95%-CI**
ERS class vs. ORS class (reference class)												
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									-0.11*	0.05	0.89	[0.81; 0.99]
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]

	11 categories				6 categories				4 categories			
	B	**(SE)**	***e^b***	**95%-CI**	***B***	**(SE)**	***e^b***	**95%-CI**	***B***	**(SE)**	***e^b***	**95%-CI**
Satisfaction with social aspects									0.30*	0.12	1.35	[1.07; 1.70]
ERS class vs. non-ERS class (reference class)												
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 to ambiguity									-0.39*	0.16	0.68	[0.49; 0.94]
Job position (level 5)	0.51*	0.25	1.66	[1.02; 2.71]								
Part-time occupation					0.56**	0.20	1.75	[1.18; 2.59]				
Organization size (small)	0.51**	0.18	1.67	[1.18; 2.35]								
Organization size (middle)					-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]								
Non-ERS class vs. ORS class (reference class)												

	11 categories				**6 categories**				**4 categories**			
	B	**(SE)**	e^b	**95%-CI**	***B***	**(SE)**	e^b	**95%-CI**	***B***	**(SE)**	e^b	**95%-CI**
Constant	2.32*	1.01										
Inattentive responding									0.75*	0.32	2.11	[1.12; 3.99]
Invariant responding	0.86**	0.28	2.37	[1.37; 4.11]	0.87**	0.29	2.38	[1.35; 4.19]	1.58***	0.38	4.84	[2.31; 10.14]
Random responding	-1.38**	0.42	0.25	[0.11; 0.57]					-1.61*	0.69	0.20	[0.05; 0.78]
Neuroticism					-0.19*	0.09	0.83	[0.70; 0.98]				
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-R^2 (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. ERS class: latent class with extreme response style, ORS class: latent class with ordinary response style, non-ERS class: latent class with 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 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 with a gray background.

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

responses (for the 6-category condition). In the 4-category condition, both latent classes could be additionally separated by the low tolerance of 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 response format, the effects of age, job relevance, and extraversion could not be distinguished among the latent classes by controlling for other predictors.

4.8 Discussion

In this paper, using a randomized experimental design, we investigated whether the shortening of an 11-point rating scale to a 6- or 4-point rating scale would reduce the presence of RSs exhibited in assessing different aspects of 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, which 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 a methodological nuisance but should instead be considered a trait of substantive meaning (Kieruj & Moors, 2013). In particular, this refers to the ERS, which 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 & Moors, 2013; Weijters et al., 2010a; 2010b; Wetzel et al., 2013; Wetzel et al., 2016; Zettler et al., 2015). As a result, our study seems to indicate that some RSs can hardly be eliminated by optimizing 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 a reduced presence of RSs and are therefore more optimal than the long rating scale. It was found that the proportion of respondents exhibiting inappropriate category use (the non-ERS class) decreased with the reduction in the number of response categories (from 23% to 16% and 12% in the 11-, 6-, and 4-category conditions, respectively). Coincidently, 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). This finding suggests that respondents can effectively cope with a less demanding rating scale. Therefore, shortening a rating scale may reduce the number of respondents using RSs as an adjustment strategy due to the inadequate length of the rating scale. This finding is in accordance with the Krosnick's concept (1991) 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 sub-optimal 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 response format, many response categories were ignored. As such, offering short rating scales may avoid overloading the respondents' differentiation ability. This finding is consistent with the recommendation provided by previous research that an optimal rating scale length for the general population should not exceed six or seven response categories (e.g., Lozano et al., 2008). These conclusion can be supported by the findings on reliability, which 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 (Bandolos & 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, researchers may be able to prevent the use of RSs as a form of adjustment strategy and thus obtain high-quality data. Generally speaking, we can recommend both short rating scales for data collection. In this study, these rating scales exhibited few differences in terms of category use, but it is worth noticing that a 6-point rating scale may allow researchers to obtain more reliable 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 what respondent characteristics and job-related factors would consistently explain the RSs that were found under the conditions, in which JS was measured with the different number of response categories. 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 aspects of 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 to ambiguity), cognitive ability, types of careless responding and the majority of job-related factors (organization size, part-time working, job skills, and tenure at current position). This finding indicates that the characteristics of respondents using a certain RS as an adjustment strategy may differ depending on response format length, whereas the respondents who use RSs as a form of individual response style have consistent personality profiles. Age, job relevance, and extraversion were found to be statistically significant in none of the experimental conditions.

4.8.1 Limitations 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 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 (Cabooter 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 remain unclear (Hamby & Levine, 2016). For this reason, the 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 (Cabooter 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 (6- and 4-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 & 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 Cabooter, Weijters, De Beuckelaer, & Davidov, 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).

A further suggestion for future research is to examine the antecedents of inappropriate responses. The present study studied 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.

4.9 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. doi: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. doi: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. doi: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. doi: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. doi:10.1177/0049124107313901

Borgers, N., Hox, J., & Sikkel, 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. doi: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. doi:10.1348/096317909x478557

Bozdogan, H. (1994). Mixture-model cluster analysis using model selection criteria and a new informational measure of complexity. In H. Bozdogan (Eds.), *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. doi:10.1177/1745691610393980

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. doi:10.1007/s11135-016-0411-5

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. doi:10.1016/j.jbusres.2015.10.138

Cacioppo, J. T, & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology, 42*(1), 116-131. doi:10.1037/0022-3514.42.1.116

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. doi: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, G. A., Jr., & 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. doi: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. doi: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. doi: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. doi: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. doi: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. doi: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. doi:10.1080/1091367X.2013.807265

Gerber-Braun, B. (2010). *The Double Cross: Individual differences between respondents with different response sets and styles on questionnaires.* Dissertation, Ludwig–Maximilians–Universität, München.

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. doi:10.2307/3172568

Greenleaf, E. A. (1992b). Measuring extreme response style. *Public Opinion Quarterly*, *56*(3), 328-351. doi:10.1086/269326

Hamby, T., & Levine, D. S. (2016). Response-scale formats and psychological distances between categories. *Applied Psychological Measurement*, *40*(1), 73-75. doi:10.1177/0146621615597961

Harzing, A. W., Baldueza, 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. doi: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. doi: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*. doi: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. doi: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. doi: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. doi: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. doi: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. doi: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. doi: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. doi: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. doi: 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. doi: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. doi:10.1016/j.jrp.2013.01.014

Kutscher, T., Crayen, C., & Eid, M. (submitted). Required Sample Size and Model Selection for Mixed Polytomous Item Response Models applied on short-scale data assessed with many response categories.

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*. doi:10.3389/fpsyg.

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. doi:10.1177/0734282914522200

Lozano, L. M., García-Cueto, E., and Muñiz, J. (2008). Effect of the number of response categories on the reliability and validity of rating scales. *Methodology*, *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. doi: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. doi: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. doi: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. Shigemasu, A. Okada, T. Imaizumi, & T. Hoshino (Eds.), *New trends in psychometrics* (pp. 253–262). Tokyo, Japan: Universal Academy Press.

Maydeu-Olivares, A., Kramp, U., García-Forero, C., Gallardo-Pujol, D., and 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. doi: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. doi: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. doi: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. doi: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. doi: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.

Naemi, B. D., Beal, D. J., & Payne, S. C. (2009). Personality predictors of extreme response style. *Journal of Personality*, *77*(1), 261-286. doi: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. doi:10.1080/10705510701575396

O'Muircheartaigh, C., Krosnick, J. A., & Helic, A. (1999, May). *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. doi:10.3758/s13428-013-0434-y

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. doi: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. doi: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. doi: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. doi: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. doi: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.* John Wiley & Sons.

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: UK: NFER-N.

Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika, 52*(3), 333–343. doi:10.1007/BF02294360

Shaftel, J., Nash, B. L., & Gillmor, S. C. (2012, April). 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, CA US: 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. doi:10.1177/0049124112452527

Summerfield, M., Bevitt, 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. doi: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). Springer, Cham.

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., Couper, M.P., & Conrad, F. (2007). Colors, labels, and interpretive heuristics for response scales. *Public Opinion Quarterly, 71*(1), 91–112. doi:10.1093/poq/nfl046

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response.* New York: Cambridge University Press.

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. doi:10.1093/ijpor/eds021

Vermunt, J.K., & Magidson, J. (2013). *Technical Guide for Latent GOLD 5.0: Basic, Advanced, and Syntax*. Statistical Innovations Inc, Belmont.

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. doi: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. doi: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. doi: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. doi: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. doi:10.1177/0146621609338593

Weijters, B., Geuens, M., & Schillewaert, N. (2010b). The stability of individual response styles. *Psychological Methods, 15*(1), 96-110. doi: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. doi: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. doi: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. doi: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. doi:10.1177/1073191115583714

Winkler, N., Kroh, M., & Spiess, M. (2006). Entwicklung einer deutschen Kurzskala 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

Zettler, I., Lang, J. W., Hülsheger, 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. doi:10.1111/jopy.12172

4.10 Appendix to Chapter 4

4.10.1 Dimensional structure for the short versions of established measures and the job satisfaction measure

Table 4.7. Standardized factor loadings of the confirmatory factor analysis for the short version of the tolerance to ambiguity scale.

Item	**Factor 1**
1. A problem has little attraction for me if I don't think it has a solution.	0.53
2. I am just a little uncomfortable with people unless I feel that I can understand their behavior.	0.49
3. It bothers me when I don't know how other people react to me.	0.48
4. If I were a scientist, it would bother me that my work would never be completed (because science will always make new discoveries).	0.46
5. Before an examination, I feel much less anxious if I know how many questions there will be.	0.36
6. I don't like to work on a problem unless there is a possibility of coming out with a clear-cut and unambiguous answer.	0.67

Notes. All items are recoded.

$\chi^2(9) = 366.97$, $p < .001$; CFI = .94; TLI = 0.90; RMSEA = .08, 90%-CI [0.07; 0.08]; SRMR = 0.05.

All factor loadings are significant on the level of $p < .001$.

These and further identical analyses were conducted using the R package lavaan (Rosseel, 2012).

Table 4.8. Standardized factor loadings of the confirmatory factor analysis for the short version of the need for cognition scale.

	CFA
Item	**Factor 1**
1. I would prefer complex to simple problems.	0.60
2. I like to have the responsibility of handling a situation that requires a lot of thinking.	0.71
3. Thinking is not my idea of fun. (recoded)	0.76
4. I would rather do something that requires little thought than something that is sure to challenge my thinking abilities. (recoded)	0.76
5. I try to anticipate and avoid situations where there is likely a chance I will have to think in depth about something. (recoded)	0.71
6. I find satisfaction in deliberating hard and for long hours.	0.60
7. I only think as hard as I have to. (recoded)	0.64
8. I like tasks that require little thought once I've learned them. (recoded)	0.57
9. The idea of relying on thought to make my way to the top appeals to me.	0.58
10. Learning new ways to think doesn't excite me very much. (recoded)	0.67

Notes. χ^2 (35) = 862.07, p < .001; CFI = .98; TLI = 0.98; RMSEA = .06, 90%-CI [0.055; 0.062]; SRMR = 0.05.

All factor loadings are significant on the level of p < .001.

Table 4.9. Results of the confirmatory factor analysis for the job characteristics measure.

Item	1. Job autonomy	2. Job skills	3. Job-related stress	4. Job security
1. I have a lot of freedom to decide how I do my own work.	0.83			
2. I have a lot of say about what happens on my job.	0.79			
3. I have a lot of freedom to decide when I do my work.	0.73			
4. My job often requires me to learn new skills.		0.70		
5. I use many of my skills and abilities in my current job.		0.80		
6. My job is more stressful than I had ever imagined.			0.73	
7. I fear that the amount of stress in my job will make me physically ill.			0.92	
8. I have a secure future in my job.				0.99
9. The company I work for will still be in business 5 years from now.				0.45
10. I worry about the future of my job. (recoded)				0.45
1. Job autonomy	-			
2. Job skills	0.65	-		
3. Job-related stress	-0.20	0.10		
4. Job security	0.35	0.42	-0.20	-

Notes. The table contains only standardized factor loadings.

χ^2 (29) = 1665.85, p < .001; CFI = .93; TLI = 0.90; RMSEA = .09, 90%-CI [0.09; 0.09]; SRMR = 0.07.

All factor loadings and intercorrelation coefficients are significant on the level of p < .001.

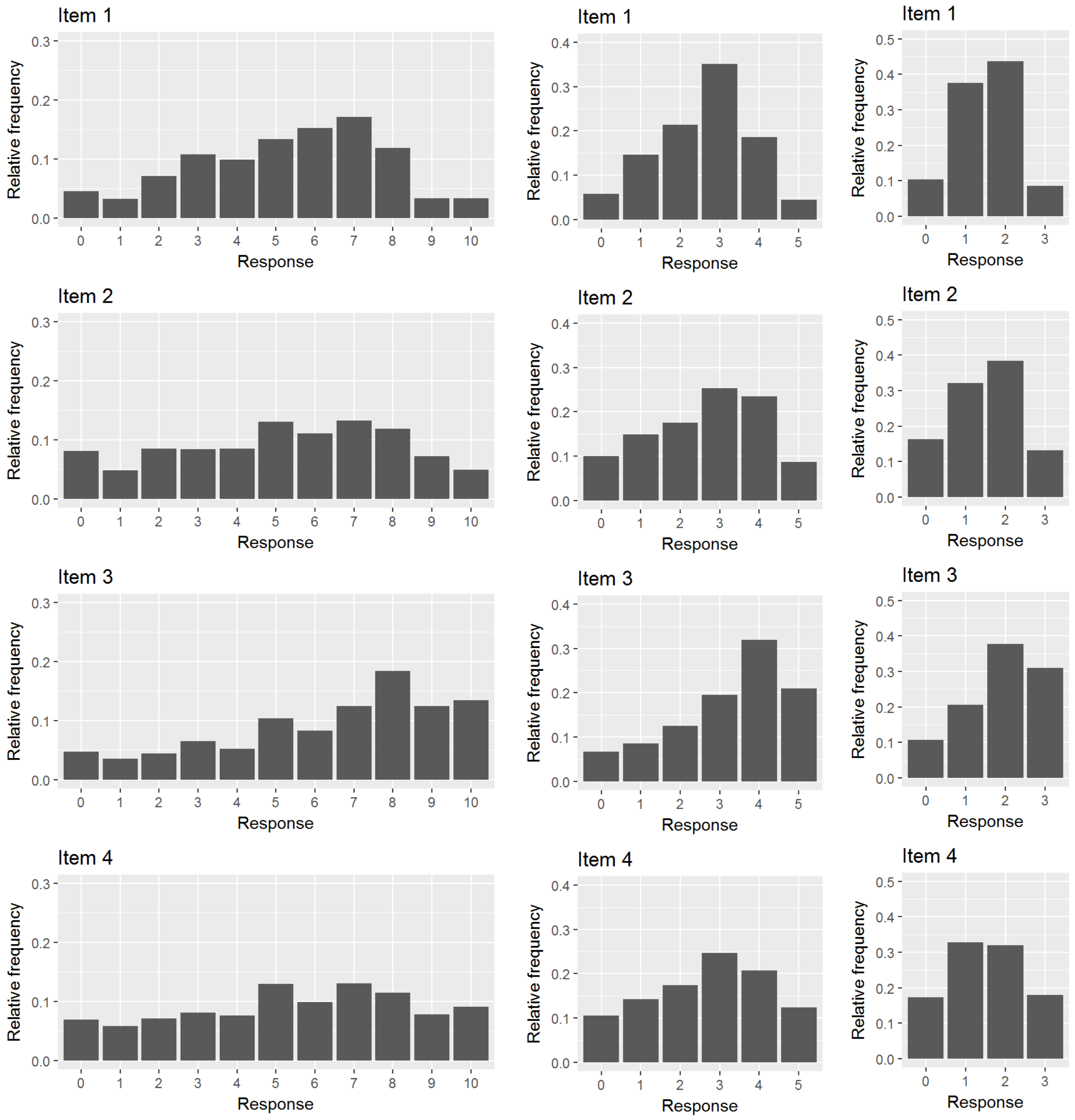
Table 4.10. Results of the exploratory factor analysis for the job satisfaction items in the three experimental conditions.

	11 categories			6 categories			4 categories		
Item	**Factor 1**	**Factor 2**	**Factor 3**	**Factor 1**	**Factor 2**	**Factor 3**	**Factor 1**	**Factor 2**	**Factor 3**
1. Your total pay			0.83		0.78				0.80
2. The monetary and non-monetary fringe benefits			0.94		0.90				0.85
3. Your job security			0.41		0.44				0.37
4. The career opportunities			0.44	0.32	0.42				0.45
5. The work tasks themselves		0.40				0.36		0.43	
6. The hours you work		0.90				0.91		0.86	
7. The work conditions		0.83		0.32		0.58		0.75	
8. The flexibility to balance work and non-work commitments		0.71				0.68		0.64	
9. The relation and communication with your immediate supervisor	0.91			0.90			0.86		
10. Appreciation, recognition, and rewards for your good work	0.97			0.89			0.89		
11. The work atmosphere and relations with your co-workers	0.70			0.83			0.69		
12. Internal staff rules and regulations in your organization	0.69			0.70			0.57		
Eigenvalues of factors	2.95	2.28	1.92	3.20	1.87	1.86	2.63	2.12	1.78
Correlations between the factors									
1. Benefits and prospects	-			-			-		
2. Work tasks and conditions	0.56	-		0.59	-		0.57	-	
3. Social aspects	0.76	0.61	-	0.74	0.59	-	0.76	0.60	-

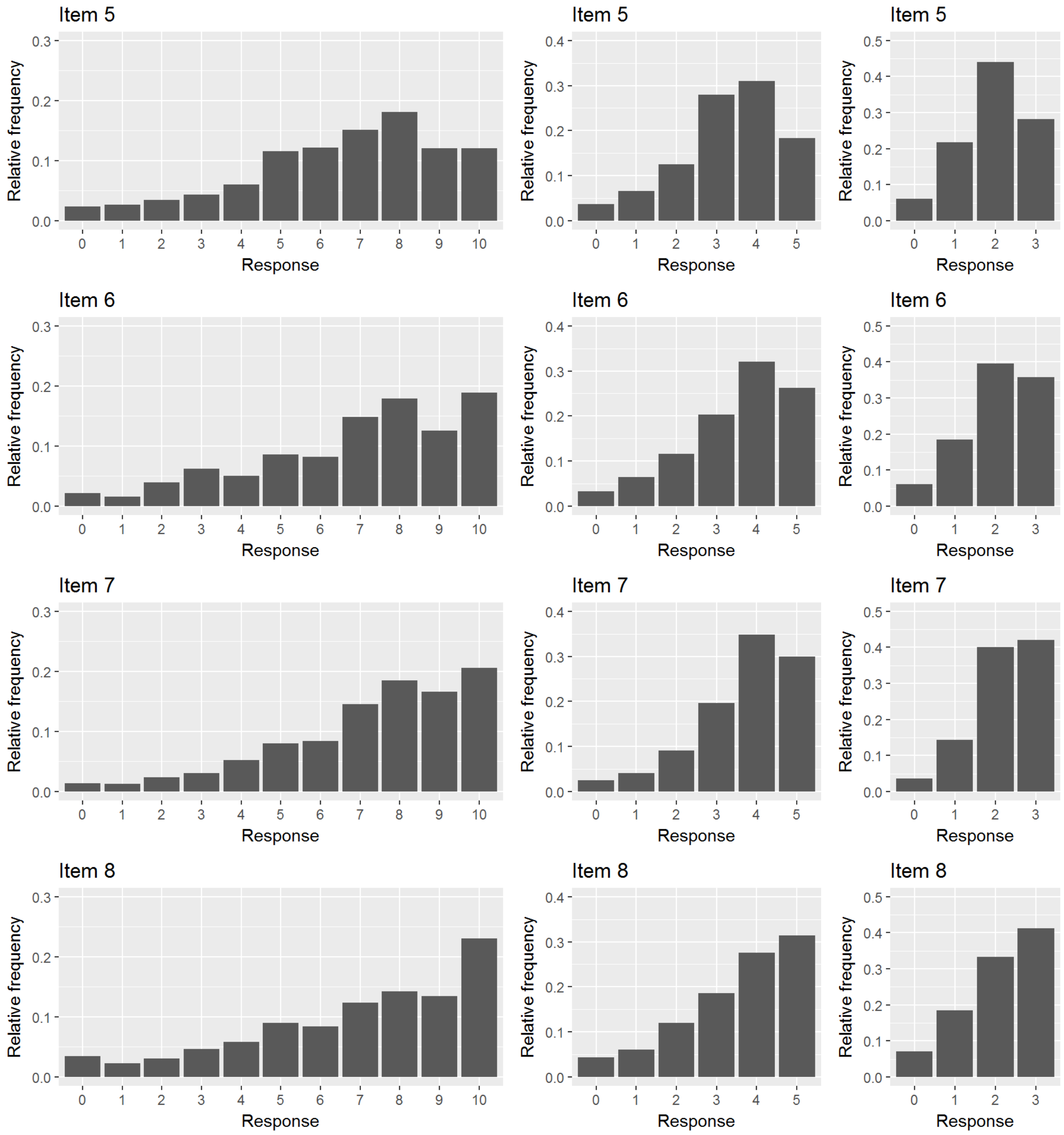
Note. The table contains only standardized factor loadings larger than |.30| (the upper part of the table).

4.10.2 Observed Category Frequencies for the Job Satisfaction Items in the three Experimental Conditions

“*Satisfaction with benefits and prospects*” subscale



"Satisfaction with work tasks and conditions" subscale



"Satisfaction with social aspects" subscale

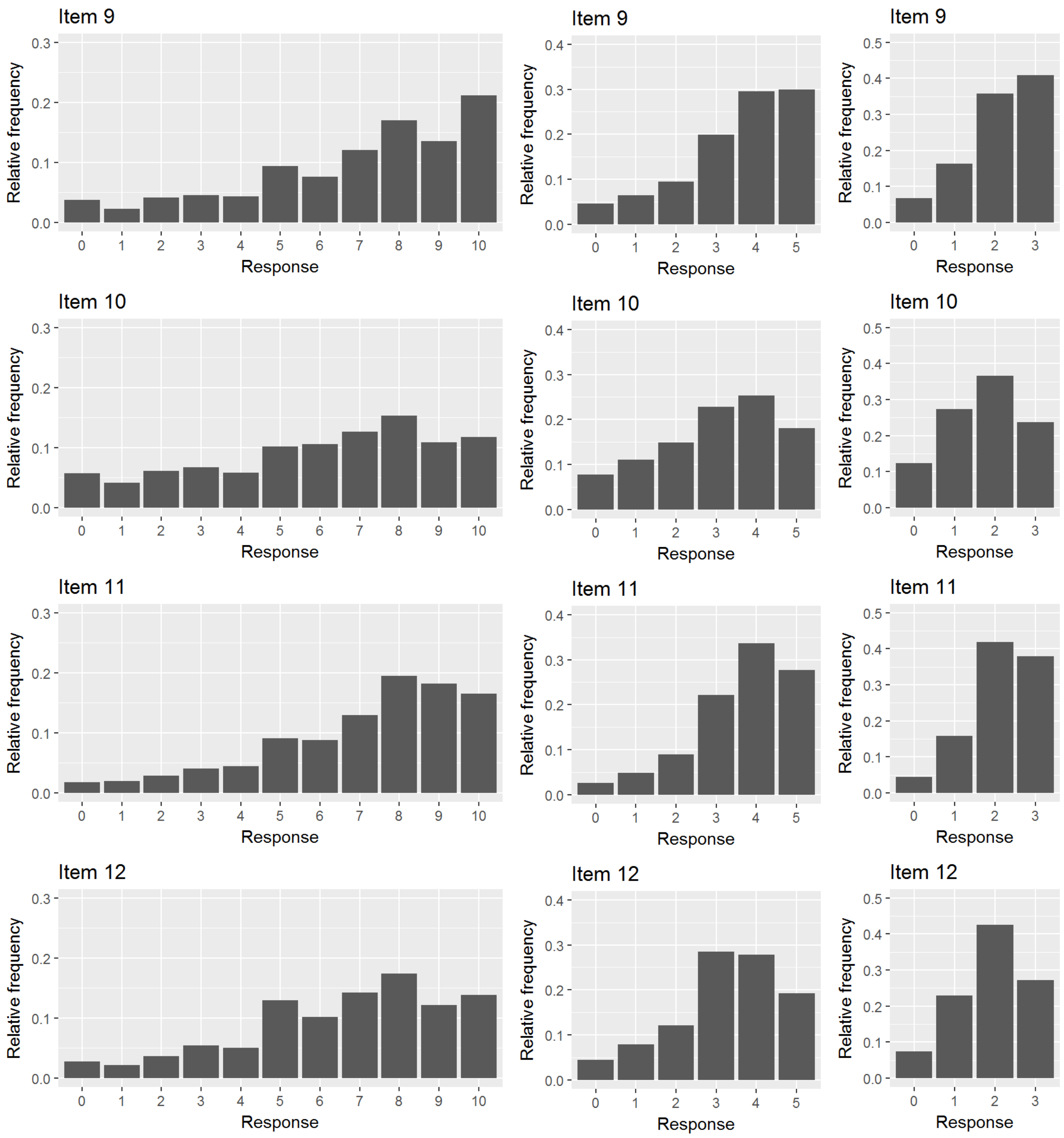


Figure 4.4. Observed category frequencies for the job satisfaction items in the three experimental conditions. (The 11-, 6-, and 4-point rating scales are in the left, middle, and right column, respectively.)

4.10.3 Latent GOLD Syntax for the three Class-Multidimensional rmGPCM

Model estimation occurred in the regression submodule using a long-formatted data file.

```
options
      algorithm
                  tolerance=1e-008 emtolerance=0.01 emiterations=8000
nriterations=600;
      startvalues
             seed=0 sets=100 tolerance=1e-005 iterations=100;
      bayes
            categorical=1 variances=1 latent=1 poisson=0;
      quadrature
            nodes=10;
      missing
            includeall;

output
      parameters=first standarderrors=robust classification
      profile frequencies  predictionstatistics estimatedvalues=model
      iterationdetails
      outfile 'class assignment and latent trait values.csv'
               classification;

variables
      caseid CASE;
      dependent   benefit, work, social;
      independent ITEM_benefit nominal,
                  ITEM_work nominal,
                  ITEM_social nominal;
      Latent      theta_benefits continuous,
                  theta_work continuous,
                  theta_social continuous,
                  class nominal 3;

equations
      class <- 1;

      theta_benefit | class;
      theta_work    | class;
      theta_social  | class;

      theta_benefit <-> theta_work;
      theta_work    <-> theta_social;
      theta_social  <-> theta_social;

      benefit <- (~diff) 1|class ITEM_benefit +
                                          (a)theta_benefit|ITEM_benefit;
      work <- (~diff) 1|class ITEM_work +
                                          (b)theta_work|ITEM_work;
      social <- (~diff) 1|class ITEM_social +
                                          (c)theta_social|ITEM_social;

a[1]=1;
b[1]=1;
c[1]=1;
```

4.10.4 Further Results from the Applications of the three Class-Multidimensional rmGPCM in the Experimental Conditions

Table 4.11. Class-specific item parameters of the three-class solution of the multidimensional rmGPCM (the 11-category condition).

	δ_i	τ_{i1tg}	τ_{i2tg}	τ_{i3tg}	τ_{i4tg}	τ_{i5tg}	τ_{i6tg}	τ_{i7tg}	τ_{i8tg}	τ_{i9tg}	τ_{i10tg}
						Class 1 (π_1 = .49)					
Item 1	1.00	-2.29	-1.73	-0.93	-0.17	-0.37	-0.22	-0.10	0.75	1.78	1.85
	(-)	(0.48)	(0.24)	(0.18)	(0.16)	(0.16)	(0.14)	(0.13)	(0.13)	(0.22)	(0.36)
Item 2	1.09	-1.23	-1.14	-0.43	-0.01	-0.56	0.07	-0.12	0.42	0.86	2.24
	(0.07)	(0.29)	(0.18)	(0.16)	(0.18)	(0.20)	(0.16)	(0.15)	(0.15)	(0.18)	(0.32)
Item 3	0.56	-3.06	-1.23	-0.45	0.34	-1.18	-0.58	-0.67	-0.78	0.49	1.51
	(0.06)	(0.39)	(0.21)	(0.19)	(0.23)	(0.23)	(0.19)	(0.15)	(0.13)	(0.14)	(0.19)
Item 4	0.68	-2.42	-1.03	-0.38	0.01	-0.63	0.18	-0.39	0.15	0.63	1.72
	(0.08)	(0.30)	(0.17)	(0.16)	(0.18)	(0.17)	(0.17)	(0.16)	(0.16)	(0.15)	(0.21)
Item 5	1.00	-2.99	-0.84	-0.63	-0.65	-0.59	-0.33	-0.21	-0.39	0.47	1.32
	(-)	(0.75)	(0.25)	(0.23)	(0.20)	(0.17)	(0.16)	(0.15)	(0.14)	(0.12)	(0.20)
Item 6	1.03	-2.24	-1.41	-0.34	-0.36	-0.55	-0.18	-0.96	-0.28	0.26	0.85
	(0.09)	(0.55)	(0.24)	(0.22)	(0.24)	(0.23)	(0.20)	(0.18)	(0.13)	(0.13)	(0.20)
Item 7	2.19	-5.04	-1.09	-0.55	-0.92	-0.46	-0.45	-0.47	-0.34	0.10	0.79
	(0.21)	(0.55)	(0.36)	(0.35)	(0.32)	(0.27)	(0.24)	(0.18)	(0.14)	(0.14)	(0.26)
Item 8	0.90	-1.42	-0.65	-0.69	-0.49	-0.41	-0.32	-0.62	-0.34	-0.06	0.63
	(0.07)	(0.35)	(0.23)	(0.24)	(0.21)	(0.21)	(0.19)	(0.17)	(0.14)	(0.12)	(0.19)
Item 9	1.00	-1.76	-1.42	-0.78	-0.66	-0.71	-0.43	-0.65	-0.77	0.14	1.00

	$\boldsymbol{\delta_i}$	$\boldsymbol{\tau_{i1tg}}$	$\boldsymbol{\tau_{i2tg}}$	$\boldsymbol{\tau_{i3tg}}$	$\boldsymbol{\tau_{i4tg}}$	$\boldsymbol{\tau_{i5tg}}$	$\boldsymbol{\tau_{i6tg}}$	$\boldsymbol{\tau_{i7tg}}$	$\boldsymbol{\tau_{i8tg}}$	$\boldsymbol{\tau_{i9tg}}$	$\boldsymbol{\tau_{i10tg}}$
	(-)	(0.35)	(0.27)	(0.27)	(0.26)	(0.24)	(0.22)	(0.19)	(0.16)	(0.13)	(0.19)
Item 10	1.49	-1.70	-1.15	-0.46	-0.51	-0.47	-0.46	-0.15	-0.02	0.50	1.84
	(0.11)	(0.37)	(0.24)	(0.29)	(0.27)	(0.21)	(0.19)	(0.16)	(0.15)	(0.18)	(0.30)
Item 11	1.04	-2.95	-1.39	-0.91	-0.61	-1.24	-0.59	-0.69	-0.54	-0.17	1.77
	(0.09)	(0.55)	(0.29)	(0.32)	(0.30)	(0.25)	(0.21)	(0.18)	(0.14)	(0.16)	(0.22)
Item 12	0.97	-2.96	-1.70	-0.76	-0.74	-1.02	-0.21	-0.68	-0.18	0.40	2.07
	(0.08)	(0.63)	(0.28)	(0.25)	(0.25)	(0.19)	(0.17)	(0.15)	(0.14)	(0.15)	(0.27)
Class 2 ($\boldsymbol{\pi_2}$ = .28)											
Item 1	1.00	0.58	-0.55	-0.73	0.08	-0.38	0.05	0.09	0.54	1.44	-0.41
	(-)	(0.25)	(0.28)	(0.25)	(0.21)	(0.21)	(0.20)	(0.19)	(0.20)	(0.30)	(0.30)
Item 2	1.09	0.97	-0.85	0.12	-0.37	-0.38	0.64	-0.23	0.31	0.65	-0.23
	(0.07)	(0.24)	(0.26)	(0.26)	(0.26)	(0.21)	(0.25)	(0.25)	(0.22)	(0.26)	(0.27)
Item 3	0.56	2.43	-0.18	-1.38	0.58	-2.15	1.19	-0.74	-0.90	0.28	-1.68
	(0.06)	(0.32)	(0.40)	(0.37)	(0.35)	(0.30)	(0.27)	(0.28)	(0.23)	(0.20)	(0.17)
Item 4	0.68	1.29	-0.20	-0.23	0.68	-1.84	1.08	-0.31	0.29	1.05	-2.10
	(0.08)	(0.22)	(0.26)	(0.27)	(0.30)	(0.27)	(0.24)	(0.25)	(0.25)	(0.30)	(0.28)
Item 5	1.00	0.42	-0.46	0.03	-0.70	-1.19	0.13	-0.40	-0.46	0.52	-0.97
	(-)	(0.33)	(0.38)	(0.40)	(0.40)	(0.30)	(0.26)	(0.24)	(0.19)	(0.19)	(0.19)
Item 6	1.03	1.19	-1.17	-1.30	0.38	-0.77	-0.19	-0.65	0.04	0.40	-1.87
	(0.09)	(0.46)	(0.51)	(0.34)	(0.32)	(0.34)	(0.28)	(0.25)	(0.23)	(0.27)	(0.26)
Item 7	2.19	-0.12	-0.77	-0.62	-0.22	-1.02	0.03	-0.59	-0.02	-0.11	-0.77
	(0.21)	(0.46)	(0.51)	(0.45)	(0.50)	(0.49)	(0.31)	(0.30)	(0.28)	(0.29)	(0.22)
Item 8	0.90	1.57	-0.89	-0.05	-0.92	-1.33	0.71	-0.77	-0.20	0.16	-2.34

	δ_i	τ_{i1tg}	τ_{i2tg}	τ_{i3tg}	τ_{i4tg}	τ_{i5tg}	τ_{i6tg}	τ_{i7tg}	τ_{i8tg}	τ_{i9tg}	τ_{i10tg}
	(0.07)	(0.40)	(0.48)	(0.46)	(0.44)	(0.33)	(0.32)	(0.34)	(0.28)	(0.28)	(0.24)
Item 9	1.00	0.76	-1.40	-0.05	-0.66	-1.99	0.88	-0.77	-0.54	0.76	-2.17
	(-)	(0.43)	(0.44)	(0.41)	(0.47)	(0.36)	(0.31)	(0.34)	(0.28)	(0.34)	(0.35)
Item 10	1.49	0.24	-0.68	-0.75	0.29	-1.31	0.25	-0.34	-0.13	0.58	-0.78
	(0.11)	(0.31)	(0.37)	(0.35)	(0.42)	(0.41)	(0.28)	(0.28)	(0.25)	(0.29)	(0.32)
Item 11	1.04	0.01	-1.23	-1.10	-0.45	-1.66	0.47	-1.04	-0.48	0.40	-1.26
	(0.09)	(0.41)	(0.52)	(0.41)	(0.38)	(0.31)	(0.30)	(0.29)	(0.22)	(0.24)	(0.24)
Item 12	0.97	0.51	-0.44	-1.61	-0.24	-1.88	0.71	-0.71	-0.34	0.66	-1.54
	(0.08)	(0.35)	(0.56)	(0.50)	(0.35)	(0.30)	(0.25)	(0.29)	(0.24)	(0.25)	(0.24)
Class 3 (π_3 = .23)											
Item 1	1.00	-8.87	-3.46	-1.99	-1.39	-1.01	-0.20	0.56	1.51	3.72	4.70
	(-)	(0.80)	(0.79)	(0.42)	(0.36)	(0.29)	(0.18)	(0.21)	(0.28)	(0.68)	(3.26)
Item 2	1.09	-3.37	-2.52	-1.98	-1.36	-0.93	-0.03	0.53	1.19	3.02	6.95
	(0.07)	(1.80)	(0.78)	(0.55)	(0.36)	(0.29)	(0.19)	(0.23)	(0.29)	(0.55)	(7.95)
Item 3	0.56	-13.02	-2.94	-4.57	-1.22	-1.70	0.58	-0.70	0.32	4.50	7.09
	(0.06)	(1.11)	(1.42)	(0.85)	(0.27)	(0.24)	(0.19)	(0.19)	(0.21)	(0.44)	(8.99)
Item 4	0.68	-3.64	-2.39	-2.98	-1.64	-1.20	0.15	0.05	1.55	3.18	10.08
	(0.08)	(1.00)	(0.71)	(0.56)	(0.31)	(0.23)	(0.19)	(0.19)	(0.23)	(0.42)	(6.02)
Item 5	1.00	-8.56	-2.62	-2.01	-1.09	-1.41	-0.24	-0.14	0.98	1.89	7.36
	(-)	(0.91)	(0.96)	(0.46)	(0.34)	(0.27)	(0.18)	(0.18)	(0.22)	(0.37)	(1.74)
Item 6	1.03	-1.58	-9.01	-2.82	-0.62	-1.23	-0.11	-0.17	0.28	1.92	3.74
	(0.09)	(0.64)	(0.66)	(0.57)	(0.28)	(0.27)	(0.19)	(0.21)	(0.21)	(0.38)	(1.54)
Item 7	2.19	-2.11	-5.17	-2.07	-1.27	-0.86	-0.45	-0.23	0.45	1.31	4.33

	$\boldsymbol{\delta_i}$	$\boldsymbol{\tau_{i1tg}}$	$\boldsymbol{\tau_{i2tg}}$	$\boldsymbol{\tau_{i3tg}}$	$\boldsymbol{\tau_{i4tg}}$	$\boldsymbol{\tau_{i5tg}}$	$\boldsymbol{\tau_{i6tg}}$	$\boldsymbol{\tau_{i7tg}}$	$\boldsymbol{\tau_{i8tg}}$	$\boldsymbol{\tau_{i9tg}}$	$\boldsymbol{\tau_{i10tg}}$
	(0.21)	(0.84)	(0.92)	(0.76)	(0.40)	(0.33)	(0.27)	(0.22)	(0.28)	(0.40)	(1.41)
Item 8	0.90	-8.07	-3.77	-2.46	-1.00	-1.00	-0.21	-0.18	0.52	1.56	2.62
	(0.07)	(1.26)	(1.29)	(0.45)	(0.26)	(0.23)	(0.18)	(0.17)	(0.22)	(0.31)	(0.82)
Item 9	1.00	-2.78	-4.50	-2.80	-1.08	-1.92	-0.51	-0.43	0.79	2.46	3.42
	(-)	(2.18)	(1.86)	(0.59)	(0.54)	(0.45)	(0.25)	(0.19)	(0.23)	(0.49)	(1.56)
Item 10	1.49	-7.45	-3.28	-2.53	-1.25	-1.13	-0.28	0.22	0.95	2.99	6.34
	(0.11)	(1.23)	(1.64)	(0.85)	(0.68)	(0.41)	(0.28)	(0.24)	(0.31)	(1.12)	(1.91)
Item 11	1.04	1.94	-8.50	-3.95	-2.02	-1.49	-0.75	-0.22	0.26	2.98	4.08
	(0.09)	(2.47)	(2.11)	(1.22)	(0.66)	(0.33)	(0.26)	(0.22)	(0.21)	(0.94)	(1.78)
Item 12	0.97	-2.93	-2.94	-3.28	-0.97	-2.07	-0.29	0.17	0.78	3.82	8.67
	(0.08)	(1.32)	(0.94)	(0.92)	(0.51)	(0.35)	(0.21)	(0.21)	(0.21)	(1.13)	(1.65)

Notes. Threshold parameters τ_{istg} are transformed from differences between two adjacent categories parameters $\beta_{0ixtg} - \beta_{0ix-1tg}$ obtained in Latent GOLD, as follows $\tau_{istg} = -1 * (\beta_{0ixtg} - \beta_{0ix-1tg}) / \delta_i$ (Vermunt & Magidson, 2006). Robust standard errors in brackets are calculated by Latent GOLD for the parameters $\beta_{0ixtg} - \beta_{0ix-1tg}$.

Table 4.12. Class-specific item parameters of the three-class solution of the multidimensional rmGPCM (the 6-category condition).

	δ_i	τ_{i1g}	τ_{i2g}	τ_{i3g}	τ_{i4g}	τ_{i5g}
			Class 1 (π_1 = .55)			
Item 1	1.00	-2.78	-0.97	-0.50	0.91	3.20
	(-)	(0.53)	(0.17)	(0.11)	(0.12)	(0.32)
Item 2	1.03	-2.21	-0.68	-0.40	0.34	2.21
	(0.07)	(0.45)	(0.16)	(0.11)	(0.13)	(0.21)
Item 3	0.73	-2.53	-1.21	-0.83	-0.82	1.49
	(0.06)	(0.40)	(0.17)	(0.13)	(0.13)	(0.18)
Item 4	0.86	-2.33	-0.46	-0.42	0.42	1.85
	(0.09)	(0.34)	(0.16)	(0.12)	(0.14)	(0.21)
Item 5	1.00	-2.50	-1.37	-0.95	-0.10	1.47
	(-)	(0.33)	(0.17)	(0.12)	(0.11)	(0.17)
Item 6	0.94	-2.66	-1.23	-0.93	-0.69	0.88
	(0.08)	(0.33)	(0.18)	(0.14)	(0.11)	(0.18)
Item 7	1.91	-2.40	-1.40	-0.94	-0.61	0.69
	(0.16)	(0.61)	(0.30)	(0.21)	(0.16)	(0.27)
Item 8	0.89	-2.33	-1.08	-0.58	-0.84	0.66
	(0.07)	(0.30)	(0.18)	(0.15)	(0.12)	(0.19)
Item 9	1.00	-2.59	-1.50	-1.23	-0.63	0.84
	(-)	(0.36)	(0.23)	(0.18)	(0.14)	(0.20)
Item 10	1.26	-2.52	-1.13	-0.65	0.14	1.69
	(0.09)	(0.36)	(0.20)	(0.15)	(0.14)	(0.26)
Item 11	1.15	-3.64	-1.73	-1.52	-0.65	1.19
	(0.09)	(0.53)	(0.27)	(0.22)	(0.14)	(0.22)
Item 12	1.05	-3.61	-1.40	-1.30	0.12	2.10
	(0.08)	(0.51)	(0.19)	(0.15)	(0.14)	(0.39)
			Class 2 (π_2 = .28)			
Item 1	1.00	-0.97	-0.36	-0.43	1.53	0.90
	(-)	(0.23)	(0.18)	(0.16)	(0.24)	(0.34)
Item 2	1.03	-0.19	0.08	-0.30	0.38	0.74
	(0.07)	(0.20)	(0.22)	(0.21)	(0.21)	(0.30)
Item 3	0.73	0.24	-0.55	-0.77	-0.59	-0.59
	(0.06)	(0.23)	(0.26)	(0.25)	(0.20)	(0.19)
Item 4	0.86	0.17	-0.14	-0.35	0.57	-0.16
	(0.09)	(0.22)	(0.21)	(0.20)	(0.28)	(0.28)
Item 5	1.00	-0.63	-0.70	-1.11	-0.07	-0.33
	(-)	(0.28)	(0.28)	(0.22)	(0.20)	(0.24)
Item 6	0.94	-0.65	-0.91	-0.84	-0.35	-0.80
	(0.08)	(0.30)	(0.26)	(0.22)	(0.22)	(0.23)
Item 7	1.91	-0.75	-1.16	-0.83	-0.36	-0.60
	(0.16)	(0.44)	(0.35)	(0.28)	(0.32)	(0.30)
Item 8	0.89	0.60	-1.74	-0.95	-0.10	-1.80
	(0.07)	(0.39)	(0.38)	(0.26)	(0.25)	(0.28)

	δ_i	τ_{i1g}	τ_{i2g}	τ_{i3g}	τ_{i4g}	τ_{i5g}
Item 9	1.00	-0.96	-1.09	-1.25	-0.35	-1.15
	(-)	(0.28)	(0.29)	(0.28)	(0.25)	(0.27)
Item 10	1.26	-0.63	-0.46	-0.60	0.13	-0.34
	(0.09)	(0.28)	(0.28)	(0.26)	(0.28)	(0.34)
Item 11	1.15	-0.98	-1.84	-1.42	-0.39	-0.89
	(0.09)	(0.45)	(0.36)	(0.28)	(0.26)	(0.29)
Item 12	1.05	-0.70	-1.15	-0.98	0.02	-0.90
	(0.08)	(0.36)	(0.30)	(0.27)	(0.26)	(0.29)
Class 3 (π_3 = .16)						
Item 1	1.00	-4.89	-3.23	-1.33	1.70	5.54
	(-)	(4.25)	(0.86)	(0.35)	(0.36)	(0.80)
Item 2	1.03	-5.31	-3.21	-1.13	1.33	6.24
	(0.07)	(4.43)	(1.06)	(0.39)	(0.36)	(1.33)
Item 3	0.73	-6.79	-3.19	-1.77	0.23	6.15
	(0.06)	(4.63)	(0.50)	(0.31)	(0.30)	(1.03)
Item 4	0.86	-2.84	-6.15	-1.30	1.21	5.69
	(0.09)	(2.84)	(4.63)	(0.48)	(0.29)	(0.92)
Item 5	1.00	-4.11	-3.75	-2.09	0.23	3.93
	(-)	(2.55)	(1.43)	(0.44)	(0.23)	(0.76)
Item 6	0.94	-3.68	-3.61	-1.51	-0.06	4.05
	(0.08)	(1.76)	(1.02)	(0.39)	(0.21)	(0.66)
Item 7	1.91	-6.00	-3.90	-1.92	-0.04	4.76
	(0.16)	(2.63)	(2.93)	(0.99)	(0.37)	(5.72)
Item 8	0.89	-9.95	-4.23	-1.60	0.33	4.15
	(0.07)	(2.45)	(1.32)	(0.45)	(0.20)	(1.06)
Item 9	1.00	-11.91	-5.06	-2.93	-0.11	6.00
	(-)	(1.13)	(1.21)	(0.76)	(0.35)	(2.95)
Item 10	1.26	-10.80	-4.49	-1.73	0.60	9.05
	(0.09)	(1.21)	(1.26)	(0.63)	(0.41)	(1.99)
Item 11	1.15	-10.87	-5.24	-2.91	-0.10	5.34
	(0.09)	(1.45)	(1.30)	(0.97)	(0.35)	(1.87)
Item 12	1.05	-6.69	-5.08	-2.59	0.71	7.65
	(0.08)	(3.62)	(1.48)	(0.55)	(0.38)	(2.02)

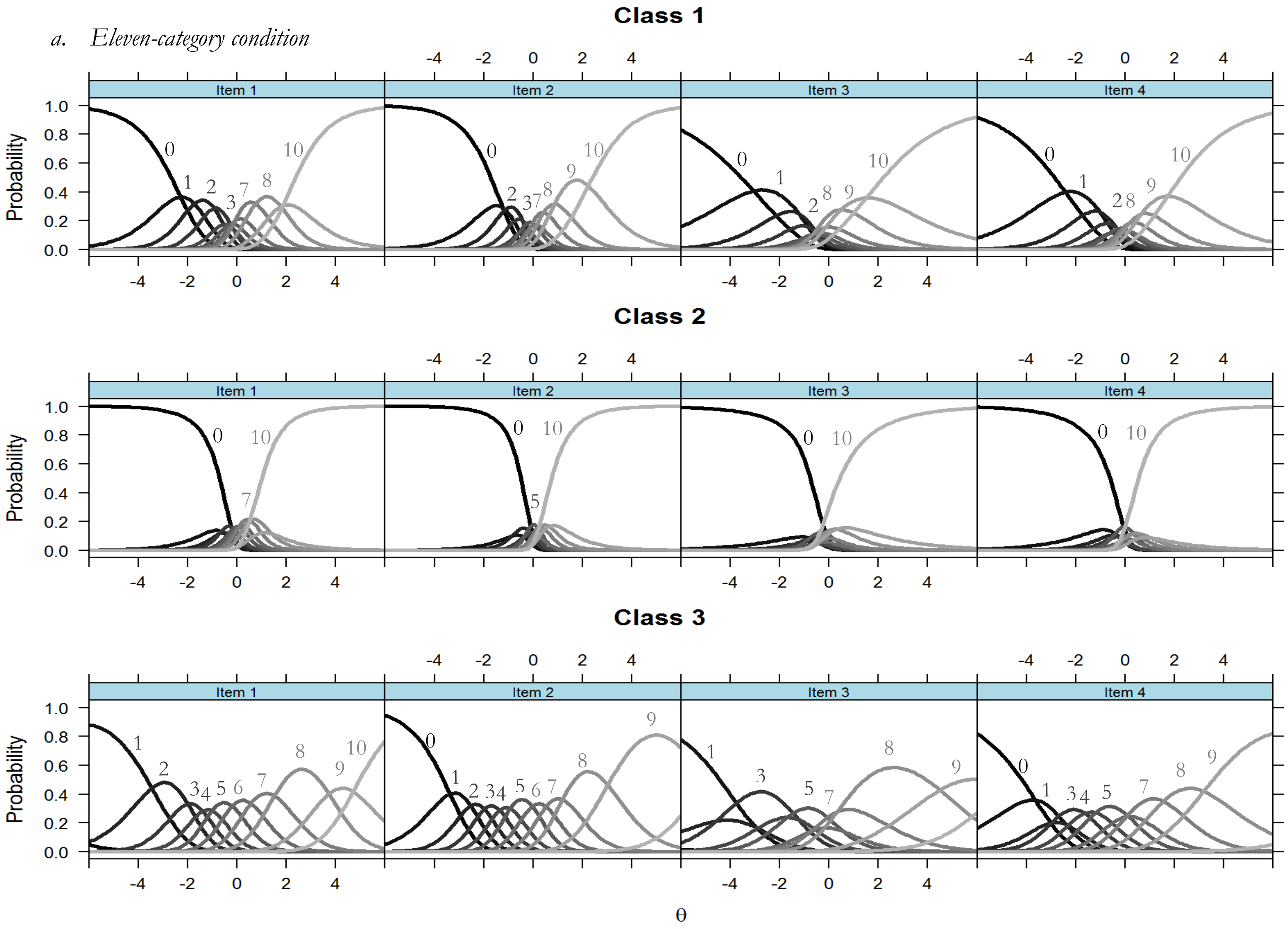
Notes. Threshold parameters τ_{istg} are transformed from differences between two adjacent categories parameters $\beta_{0ixtg} - \beta_{0ix-1tg}$ obtained in Latent GOLD, as follows $\tau_{istg} = -1 * (\beta_{0ixtg} - \beta_{0ix-1tg})/\delta_i$ (Vermunt & Magidson, 2006). Robust standard errors in brackets are calculated by Latent GOLD for the parameters $\beta_{0ixtg} - \beta_{0ix-1tg}$.

Table 4.13. Class-specific item parameters of the three-class solution of the multidimensional rmGPCM (the 4-category condition).

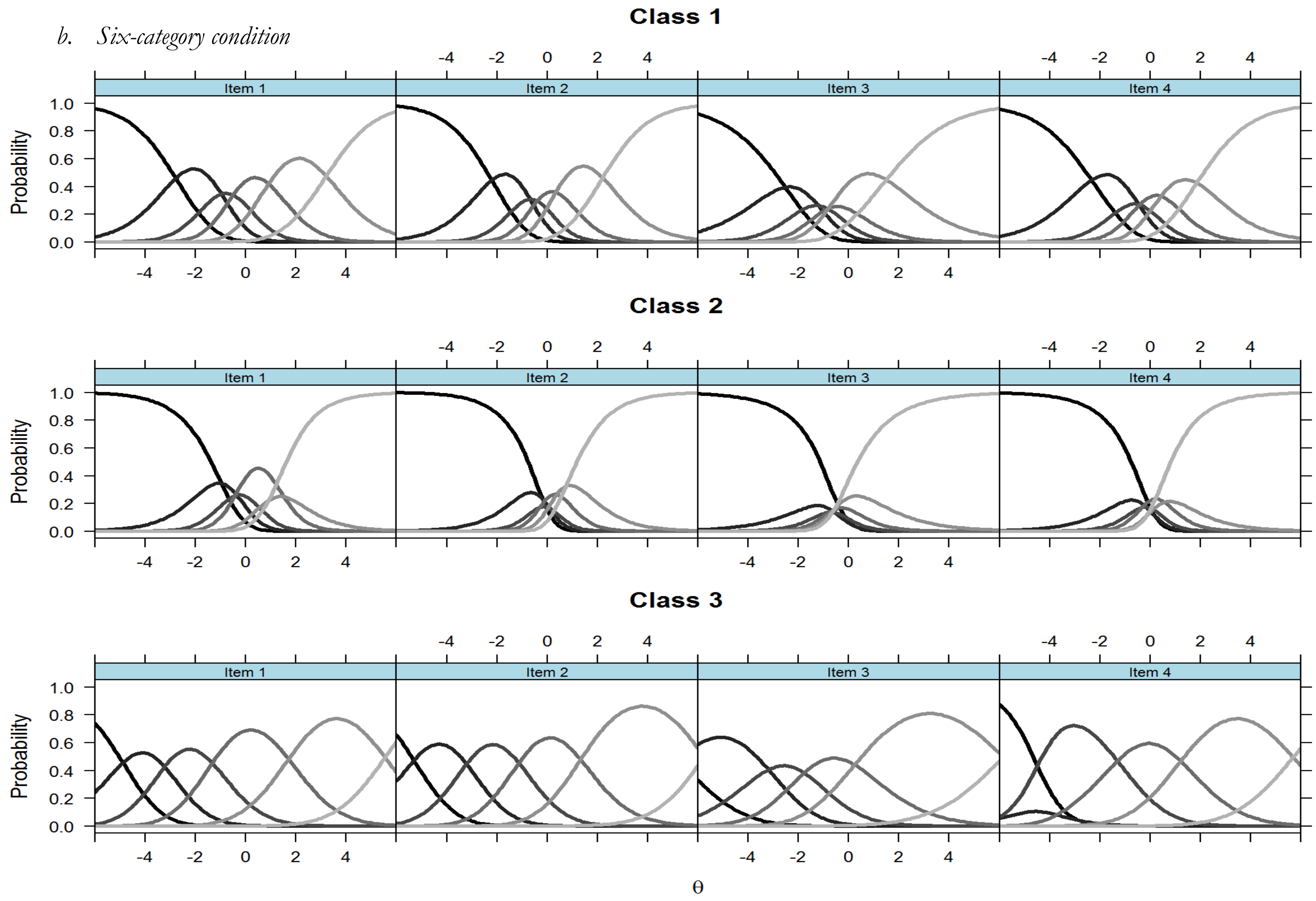
		Class 1 (π_1 = .62)			Class 1 (π_2 = .26)			Class 3 (π_3 = .12)		
	δ_i	τ_{i1g}	τ_{i2g}	τ_{i3g}	τ_{i1g}	τ_{i2g}	τ_{i3g}	τ_{i1g}	τ_{i2g}	τ_{i3g}
Item 1	1.00	-2.91	-0.25	3.15	-1.34	0.38	1.65	-5.83	-0.78	5.05
	(-)	(0.49)	(0.19)	(0.31)	(0.29)	(0.26)	(0.35)	(1.62)	(0.40)	(0.87)
Item 2	0.95	-2.12	-0.23	2.41	-0.38	-0.01	1.02	-5.18	-0.30	5.32
	(0.07)	(0.43)	(0.14)	(0.19)	(0.30)	(0.26)	(0.33)	(1.07)	(0.45)	(3.56)
Item 3	0.55	-2.46	-1.64	1.31	-0.71	-0.49	-1.58	-12.21	-1.44	3.98
	(0.06)	(0.33)	(0.12)	(0.14)	(0.26)	(0.26)	(0.23)	(9.02)	(0.30)	(0.97)
Item 4	0.83	-1.90	-0.03	1.79	-0.49	0.12	0.37	-7.01	0.15	9.55
	(0.09)	(0.35)	(0.15)	(0.37)	(0.25)	(0.31)	(0.25)	(3.99)	(0.43)	(12.34)
Item 5	1.00	-2.77	-1.14	1.08	-1.09	-0.85	-0.34	-11.43	-1.15	5.20
	(-)	(0.31)	(0.20)	(0.27)	(0.25)	(0.30)	(0.22)	(1.17)	(0.55)	(3.40)
Item 6	0.90	-2.99	-1.33	0.52	-0.86	-0.78	-0.86	-5.04	-1.71	4.61
	(0.09)	(0.42)	(0.23)	(0.42)	(0.26)	(0.26)	(0.24)	(1.07)	(0.34)	(1.32)
Item 7	1.77	-2.74	-1.36	0.37	-1.24	-1.07	-0.94	-8.44	-1.49	4.27
	(0.20)	(0.64)	(0.42)	(0.58)	(0.46)	(0.41)	(0.36)	(3.36)	(0.88)	(3.81)
Item 8	0.82	-2.59	-1.06	0.05	-0.42	-1.44	-1.28	-6.28	-1.22	4.74
	(0.07)	(0.25)	(0.15)	(0.27)	(0.32)	(0.31)	(0.18)	(3.55)	(0.32)	(2.47)
Item 9	1.00	-2.88	-1.47	0.45	-1.20	-1.14	-1.53	-10.94	-1.76	4.49
	(-)	(0.35)	(0.18)	(0.19)	(0.37)	(0.31)	(0.43)	(2.37)	(0.47)	(5.07)
Item 10	1.39	-2.19	-0.51	1.53	-0.91	-0.48	-0.12	-9.02	-0.47	4.73
	(0.12)	(0.49)	(0.20)	(0.29)	(0.32)	(0.32)	(0.51)	(1.62)	(0.68)	(14.84)
Item 11	1.00	-3.58	-1.84	0.90	-1.53	-1.32	-1.46	-11.31	-1.56	3.72
	(0.08)	(0.42)	(0.24)	(0.18)	(0.35)	(0.29)	(0.51)	(2.05)	(1.01)	(3.07)
Item 12	1.01	-3.04	-1.14	1.78	-1.07	-1.09	-1.02	-10.97	-0.68	5.88
	(0.09)	(0.39)	(0.13)	(0.25)	(0.30)	(0.29)	(0.53)	(1.39)	(0.58)	(19.71)

Notes. Threshold parameters τ_{istg} are transformed from differences between two adjacent categories parameters $\beta_{0ixtg} - \beta_{0ix-1tg}$ obtained in Latent GOLD, as follows $\tau_{istg} = -1 * (\beta_{0ixtg} - \beta_{0ix-1tg}) / \delta_i$ (Vermunt & Magidson, 2006). Robust standard errors in brackets are calculated by Latent GOLD for the parameters $\beta_{0ixtg} - \beta_{0ix-1tg}$.

a. Eleven-category condition



b. Six-category condition



c. Four-category condition

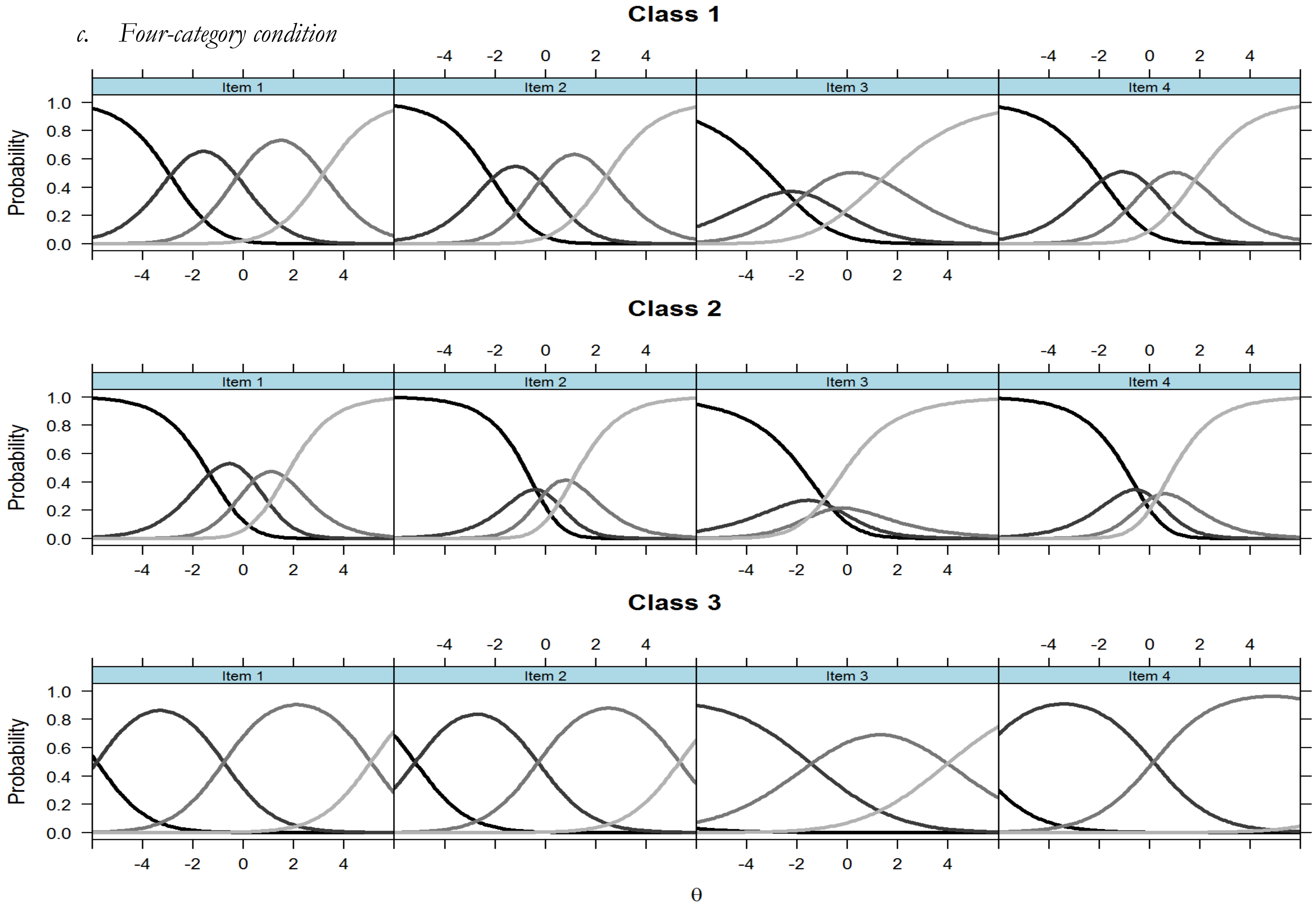
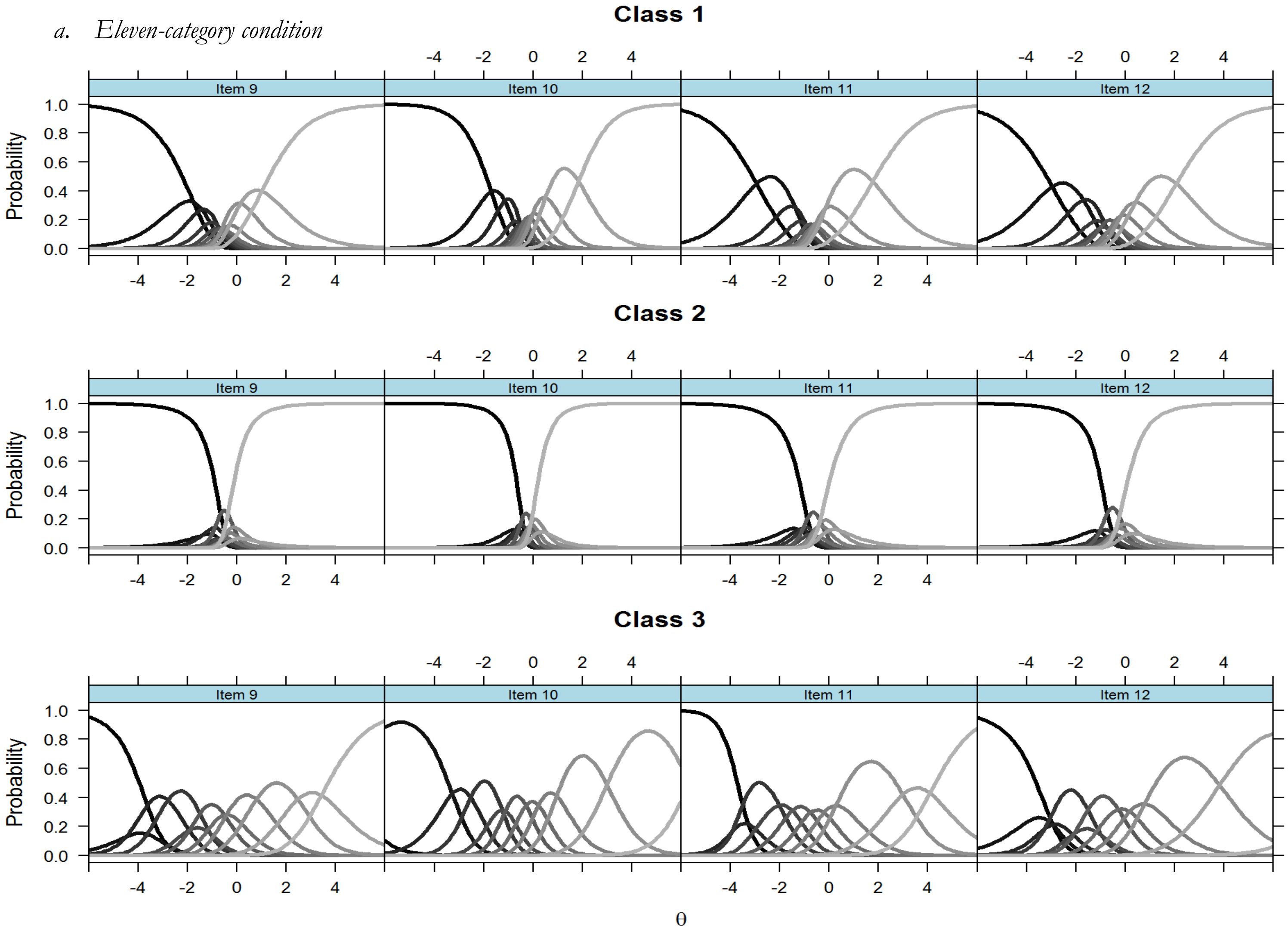


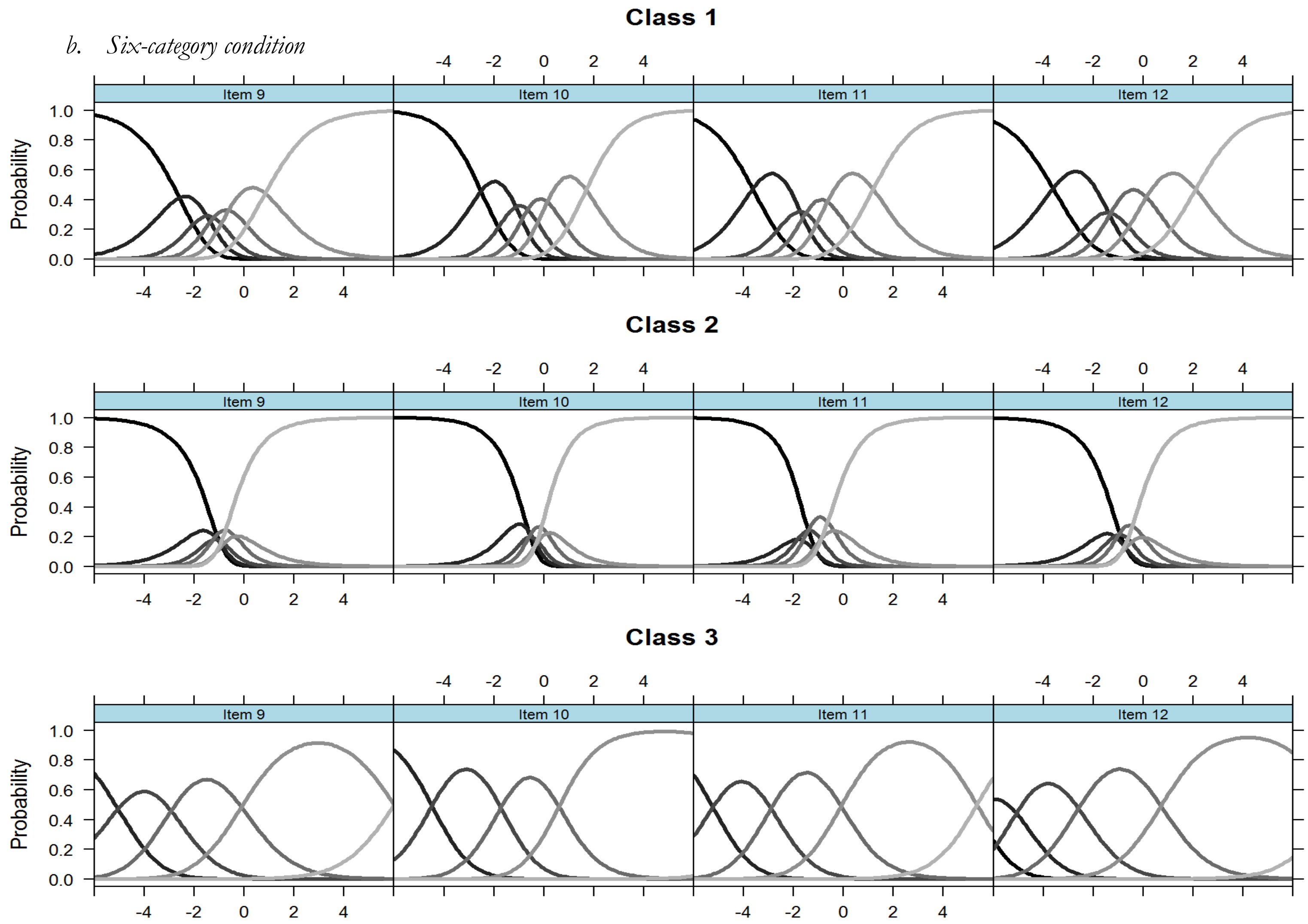
Figure 4.5. Class-specific category characteristic curves for the items of the subscale "*Satisfaction with benefits and prospects"* under the three experimental conditions.

(Categories whose response probability is the highest on a certain segment of the latent continuum are indicated with their values.)

a. Eleven-category condition



b. Six-category condition



c. Four-category condition

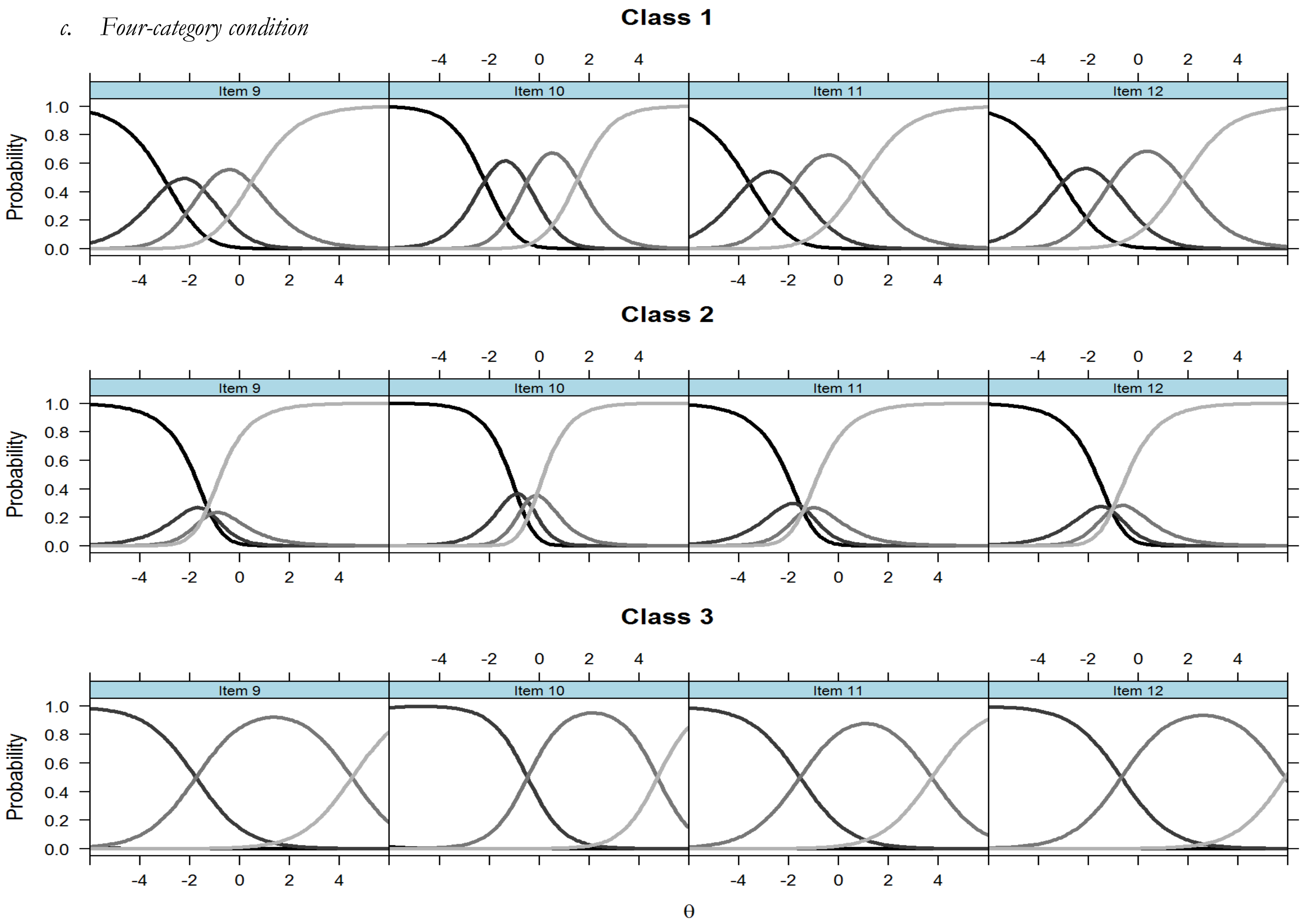


Figure 4.6. Class-specific category characteristic curves for the items of the subscale "*Satisfaction with social aspects*" under the three experimental conditions.

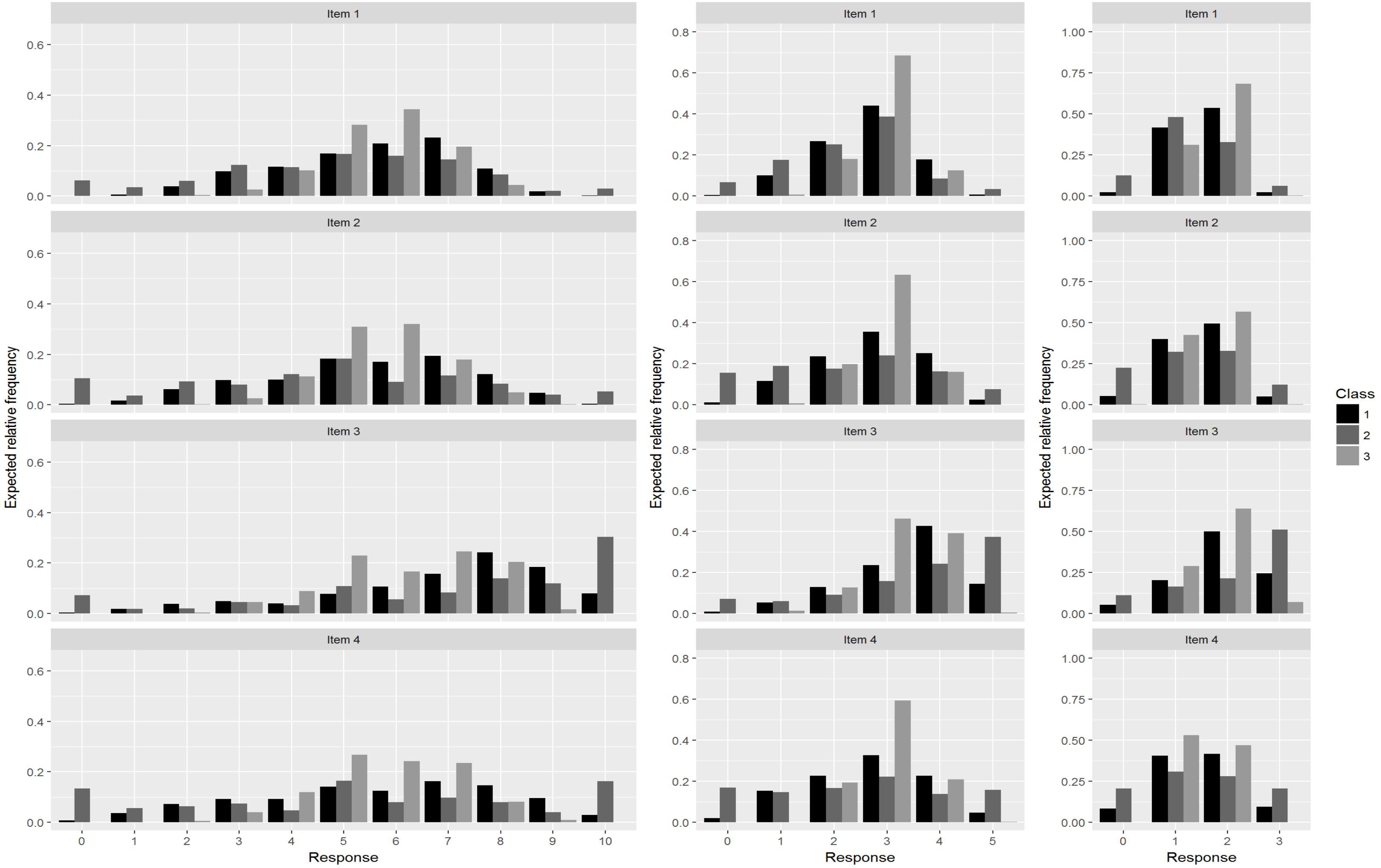


Figure 4.7. Expected relative frequencies for items of the subscale "*Satisfaction with benefits and prospects*" under the three experimental conditions. (The 11-, 6-, and 4-category conditions are in the left, middle, and right column, respectively.)

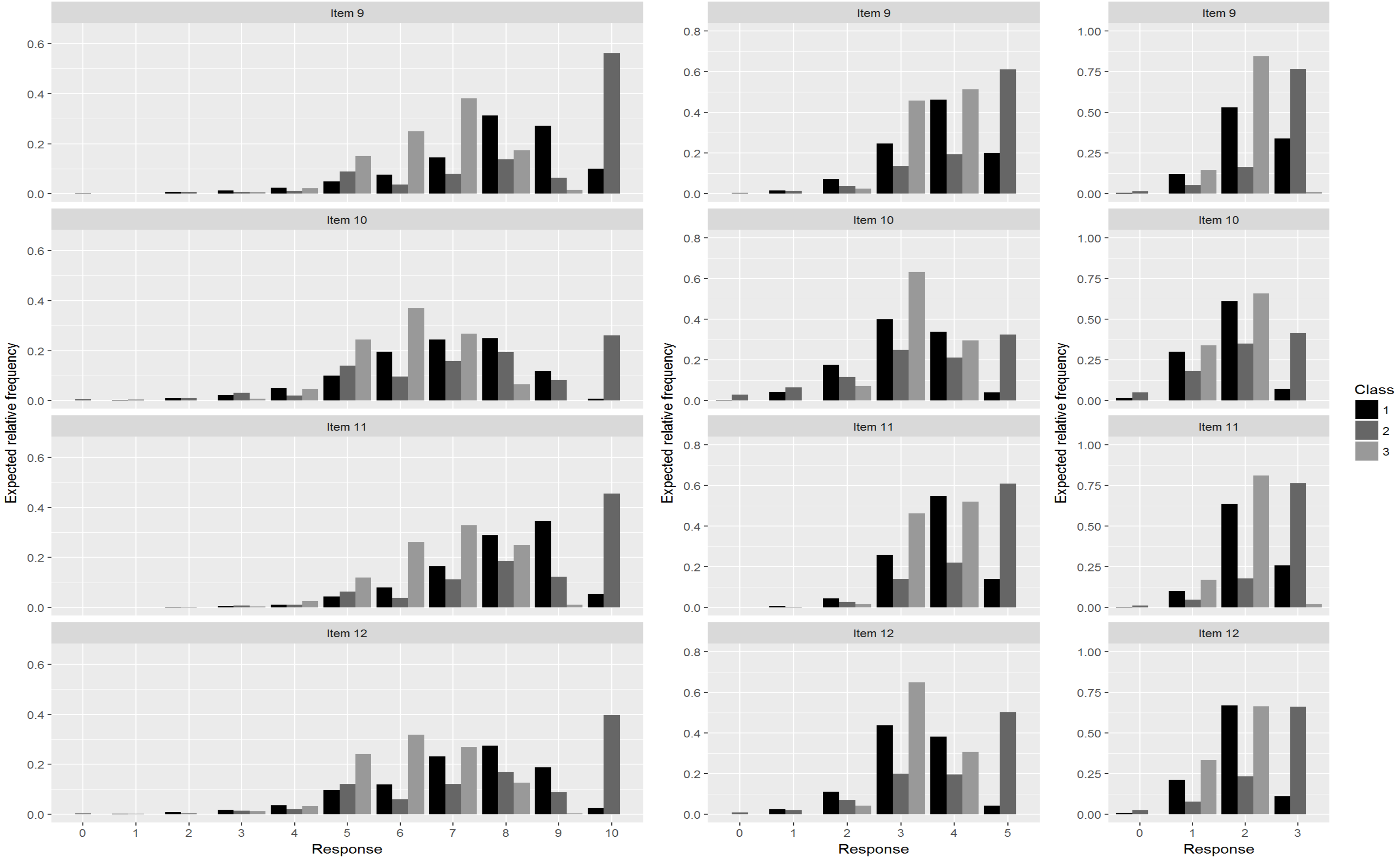


Figure 4.8. Expected relative frequencies for the items of the subscale *"Satisfaction with social aspects"* under the three experimental conditions. (The 11-, 6-, and 4-category conditions are in the left, middle, and right column, respectively.)

Table 4.14. Descriptive statistics and confidence intervals for the means of the class-specific latent trait variables (posterior distributions).

	11 categories			**6 categories**			**4 categories**		
Latent trait variable	**Mean (SD)**	**Min; Max**	**95%-CI**	**Mean (SD)**	**Min; Max**	**95%-CI**	**Mean (SD)**	**Min; Max**	**95%-CI**
Benefits and prospects									
- ORS class	0.00 (0.62)	-2.03; 1.66	[-0.04; 0.04]	-0.01 (0.84)	-2.39; 2.79	[-0.05; 0.03]	-0.01 (1.15)	-3.19; 2.68	[-0.07; 0.05]
- ERS class	0.00 (0.47)	-1.23; 1.05	[-0.04; 0.04]	0.03 (0.91)	-2.25; 1.89	[-0.05; 0.11]	0.07 (1.50)	-3.45; 2.75	[-0.05; 0.19]
- Non-ERS class	0.00 (0.89)	-2.42; 2.22	[-0.08; 0.08]	0.01 (1.47)	-5.27; 4.21	[-0.13; 0.15]	-0.08 (1.38)	-4.84; 2.85	[-0.24; 0.08]
Work tasks and conditions									
- ORS class	0.01 (0.42)	-1.34; 1.11	[-0.01; 0.03]	-0.02 (0.72)	-1.95; 1.70	[-0.06; 0.02]	-0.02 (0.94)	-2.80; 1.77	**[-0.06; 0.02]**
- ERS class	0.01 (0.40)	-1.17; 0.87	[-0.03; 0.05]	0.08 (0.78)	-2.08; 1.52	[0.02; 0.14]	0.15 (1.05)	-2.57; 1.82	**[0.07; 0.23]**
- Non-ERS class	-0.04 (0.59)	-1.77; 1.21	[-0.08; 0.00]	-0.02 (0.93)	-2.23; 1.89	[-0.12; 0.08]	-0.21 (1.21)	-3.85; 1.93	**[-0.35; -0.07]**
Social aspects									
- ORS class	0.01 (0.70)	-1.98; 1.72	[-0.03; 0.05]	-0.04 (1.05)	-3.31; 2.41	[-0.10; 0.02]	-0.06 (1.25)	-3.60; 2.45	[-0.12; 0.00]
- ERS class	0.02 (0.68)	-1.84; 1.33	[-0.04; 0.08]	-0.04 (1.05)	-3.02; 2.16	[0.01; 0.21]	0.24 (1.47)	-3.41; 2.40	**[0.12; 0.36]**
- Non-ERS class	-0.05 (0.91)	-3.23; 1.88	[-0.13; 0.03]	0.11 (1.22)	-4.41; 3.58	[-0.18; 0.14]	-0.21 (1.21)	-3.73; 2.15	[-0.35; -0.07]

Note. Non-overlapping confidence intervals on the class-specific means within a subscale are marked in bold.

5 GENERAL DISCUSSION

General Discussion

Many national panel surveys use long rating scales as the gold standard of cognitive well-being assessment. Usually, it is argued that fine response categories may reflect fine-grained differences between respondents in overall or domain-specific life satisfaction (e.g., Diener, Inglehart, & Tay, 2013; Krosnick & Presser, 2010; Preston & Colman, 2000; Willits, Theodori, & Luloff, 2016). However, recent research suggests that long rating scales have generally performed more poorly than did short scales (e.g., Freund, Tietjens, & Strauss, 2013; Hamby & Levine, 2016; Khadka, Gothwal, McAlinden, Lamoureux, & Pesudovs, 2012). Specifically, respondents who fail to differentiate correctly between a large number of fine categories are inclined to use an adjustment strategy in the form of ICU (e.g., RSs, avoiding categories, and careless responding; Baumgartner & Steenkamp, 2001; Cox, 1980; De Jong, Steenkamp, Fox, & Baumgartner, 2008; Krosnick, 1999; Miller, 1956; Shaftel, Nash, & Gillmor, 2012; Swait & Adamowicz, 2001; Viswanathan, Sudman, & Johnson, 2004; Weathers, Sharma, & Niedrich, 2005). Therefore, the presence of ICU impairs the psychometric quality of measures (e.g., Chang, 1994; Culpepper, 2013; Jin & Wang, 2014; Lee & Paek, 2014; Lozano, García-Cueto, & Muñiz, 2008; Maydeu-Olivares, Kramp, García-Forero, Gallardo-Pujol, & Coffman, 2009; Revilla, Saris, & Krosnick, 2014; Tarka, 2016). Thus, the present dissertation is part of a line of research that validates rating scale length.

In the previous chapters, three studies were presented. In Chapter 2, the adequacy of an 11-point rating scale used to assess job satisfaction in the HILDA surveys (Summerfield, Bevitt, Freidin, Hahn, La, Macalalad, ... & Wooden, 2017) was examined. Chapter 3 focused on identifying the sample size required for an effective application of the mixed polytomous IRT models to the data collected with a few items and many response categories and on examining what information criteria are suitable for accurate detection of the best-fitting class solution. In Chapter 4, with the manipulation of rating scale length, an optimal number of response categories was identified for a valid assessment of job satisfaction. Finally, both Chapters 2 and 4 provided specific characteristics of RS users. The specific research goals, methodological approaches and key findings of these three studies are summarized in Table 5.1.

This chapter begins with an integrated discussion of the central findings from these three studies under a broader perspective. Specifically, five main foci are addressed: (1) the adequacy of an 11-point rating scale, (2) an optimal rating scale length, (3) an update of the concept of ICU, (4) the personal profiles of RS users, and (5) the requirements and challenges of mixed IRT models for detecting ICU. The generalizability of these findings is then discussed. In addition, the implications of the relevant findings are presented. Lastly, suggestions for future research are made, and conclusions are drawn.

Table 5.1. Summary of the findings in the present dissertation.

Chapter	Aims	Methods	Findings
2	To examine the appropriateness of an 11-point rating scale for a valid assessment of job satisfaction.	Analysis of data from the first wave of the HILDA survey (*n* = 7,036) concerning aspects of job satisfaction (5 items assessed with an 11-point rating scale) using mixed polytomous IRT models.	The majority of the sample (60%) showed ICU (ERS class and semi-ERS class). Other respondents (40%) used an 11-point rating scale more adequately (DRS class). None of the latent classes used all 11 response categories: it was appropriately differentiated only between two to three categories in the ERS class, three to five categories in the semi-ERS class, and five to six categories in the DRS class, indicating that an 11-point rating scale is of limited adequacy for a valid assessment of job satisfaction.
3	To investigate the required sample size for the mixed polytomous IRT models when applied to data originating from panel surveys (e.g., a few items assessed with a large number of response categories). To investigate the effectiveness of information criteria for selecting the best-fitting class solution for these models.	Analysis of simulated data drawn from the population parameters obtained by an empirical application of a specific mixed polytomous IRT model (three-class solution). The sample size (*N* = 500 up to 5,000 with 500-step) and the type of the model (mPCM and rmGPCM) were manipulated.	A more parsimonious mPCM showed slightly better results than did the rmGPCM. The perfect convergence of the EM estimation algorithm was obtained in all sample-size conditions. The accuracy of parameter estimates and standard error estimates generally improved by increasing sample size. Most of the estimation problems occurred due to low category frequencies, the presence of a highly discriminating item, and small class size. Our recommendations are as follows: An insufficient sample size contains fewer than 1,500 cases; a reasonable sample size should include at least 2,500 cases. Required sample sizes for the effective performance (in terms of a high accuracy rate) of information criterion applied for identifying the best-fitting class solution: - AIC3 from *N* = 1,500 observations - SABIC from *N* = 2,500 observations - BIC and CAIC from *N* = 4,500 / 5,000 observations, respectively Insufficient accuracy was found for the AIC, regardless of the sample size.
4	To examine the optimal rating scale length for assessing job satisfaction.	Analysis of experimental data on job satisfaction (*N* = 6,999) measured in different experimental conditions with a varying number of response categories of a rating scale (4-, 6-, and 11-point rating scale). Evaluation criteria: the extent of ICU and reliability of job	Regardless of rating scale length, identical RSs were found: ORS, ERS, and non-ERS. When an 11-point rating scale was proposed, respondents ignored many categories; when one of the short rating scales was offered, few or hardly any categories were avoided. By shortening rating scale length, the proportion of respondents with ICU was reduced (from 23% to 16% and 12% with 11-, 6-, and 4-point rating scales, respectively). In contrast, the proportion of respondents who exhibited the ORS was positively associated with decreasing the number of response categories (from 49% to 55% and 62% with 11-, 6-, and 4-point rating scales, respectively). The proportion of

Chapter	Aims	Methods	Findings
		satisfaction measure. Detection of patterns of ICU using the multidimensional mixed IRT model.	respondents with the ERS was nearly the same for all rating scale lengths (28% with 11- and 6-point rating scales; 26% with a 4-point rating scale). Regarding model-based reliability scores, a 6-point rating scale was identified as the most optimal, followed by a 4-point rating scale.
2, 4	To explain RS use by means of stable respondent characteristics and contextual factors at work.	Analyses of the data using multinomial logistic regression analysis.	Two sets of predictors are identified: (i) (almost) general predictors[1] and (ii) response-format-specific predictors[2]. (Almost) general predictors of ERS use are as follows: high general self-efficacy and perceived job autonomy, high self-deception, low job-related stress, and a low or medium job position (compared with ORS use); high self-deception, neuroticism, need for cognition, and low job-related stress (compared to non-ERS use). (Almost) general predictors of non-ERS use are as follows: a low need for cognition and low job security (as opposed to ORS use). The response-format-specific predictors found are these: socio-demographic variables, cognitive ability, and most of the job-related factors.

Notes. HILDA = Household, Income and Labor Dynamics in Australia. IRT = Item response theory. ICU = Inappropriate category use. ERS = Extreme response style. Semi-ERS = Semi-extreme response style. RS = Response style. DRS = Differential response style. PCM = Partial credit model. GPCM = Generalized partial credit model. AIC 3 = Akaike's information criterion with triple the number of model parameters. SABIC = Sample-size adjusted Bayesian information criterion. BIC = Bayesian information criterion. CAIC = Consistent Akaike's information criterion. AIC = Akaike's information criterion. ORS = Ordinary response style. Non-ERS = Non-extreme response style.

[1] (Almost) general predictors include almost general predictors (which predicted the use of a certain RS under two experimental conditions) and general predictors (which predicted the use of a specific RS regardless of rating scale length).

[2] Response-format-specific predictors are those that predicted the use of a certain RS only for a particular rating scale length.

5.1 Limited Adequacy of an 11-point Rating Scale

Researchers considering the response process from the respondents' perspective were the first to scrutinize excessively large (and excessively short) rating scale lengths as have limited adequacy for a valid assessment of traits and attitudes (Cox, 1980; Krosnick, 1991; Tourangeau, Rips, & Rasinski, 2000). They emphasized that respondents are able to give valid responses about a particular trait or attitude if the response format of its measure is aligned to subjective categories in which respondents think about the item content. In other respects, respondents may experience discrepancy resulting in ICU as an adjustment strategy. Furthermore, psychometricians have confirmed these concerns by providing evidence that an enlargement of rating scale length is positively associated with an increase of ICU effects (e.g., Chang, 1994; Culpepper, 2013; Lee & Paek, 2014; Lozano et al., 2008; Maydeu-Olivares et al., 2009; Revilla et al., 2014; Tarka, 2016; Weng, 2004). However, the adequacy of an 11-point rating scale that is widely used in panel surveys for the assessment of cognitive well-being has not yet been validated.

The present dissertation supports the assumption that an 11-point rating scale has limited adequacy for the assessment of job satisfaction. In Chapter 2, we presented an empirical study examining patterns of ICU, which were detected by applying the rmGPCM. For this purpose, data on five aspects of job satisfaction assessed using an 11-point rating scale (endpoint labeling) from the representative sample of Australian employees and employers (n = 7,036; the HILDA survey) were analyzed. The analysis revealed a high presence of ICU in the responses: 60% of the sample used two types of the ERS (latent class with the traditional ERS and so-called semi-ERS class). Moreover, the class-specific response behavior could be characterized by a high number of avoided categories, indicating that the respondents needed an average of two or four categories (in ERS class and semi-ERS class, respectively) or six response categories (in DRS class). Specifically, the first two classes differentiated only between extreme categories, whereas the last class showed accurate differentiation between several lower categories and certain upper categories, and it almost did not use middle categories. This finding is consistent with evidence that persons' thinking complexity and discrimination ability are crucial factors affecting their response behaviors (Miller, 1956; Naemi, Beal, & Payne, 2009; Viswanathan et al., 2004; Weathers et al., 2005). Additional evidence for the limited adequacy of an 11-point rating scale came from the experimental study (Chapter 4). In the 11-category condition of this study, 12 items of job satisfaction were answered by American MTurk workers ($N_{11\text{-cat.}}$ = 2,322) who were randomly assigned to one of three experimental conditions. American MTurk workers represented an online sample, which generally provides high-quality data and is properly representative of the general population across several psychological traits (McCredie & Morey, 2018). By applying a multidimensional version of the rmGPCM due to the three-dimensional structure of job satisfaction items, we found another constellation of latent classes (ERS class, non-ERS class, and ORS class[11]). Interestingly, in this sample, only a slightly lower proportion of respondents used response categories inappropriately (51% in both ERS class and non-ERS class). Identical to the findings from the study reported in Chapter 2, latent classes strongly differed in the number of response categories used to respond to items: on average, two extreme categories in the ERS class, six categories (three lower and three higher categories) in the ORS class, and up to seven or eight categories (with the exception of extreme and some middle categories) in the non-ERS class, suggesting that some respondents in this sample could better tolerate a long rating scale. These differences in the response behavior between the HILDA sample and American MTurk sample may be explained by different sample characteristics: for instance, the MTurk workers have much more practice answering online questionnaires than do respondents in the HILDA sample. However, these two samples originated from English-speaking countries, so any differences in ICU due to language differences can be excluded. Nevertheless, based on the findings from these two studies, a high rate of ICU and a high number of unnecessary categories indicate that an 11-point rating scale may produce a high amount of ICU effects that could be at least partly eliminated by optimizing rating scale length. These findings suggest that a rating scale with no more than six (or eight) categories (without the middle category) is more optimal for assessing job satisfaction.

The primary significance of this research is threefold. First, the present studies (Chapters 2 and 4) are the first to illustrate the respondents' response behavior at the item and category level to evaluate the adequacy of a long rating scale for providing high-quality data. Researchers should weigh the advantages and disadvantages of assessing job satisfaction with many categories. Second, our findings allow researchers to gain essential suggestions for improving the original response format. Third, our findings were derived using the mixture polytomous IRT approach, which avoids the need to define types of ICU a priori. Moreover, in contrast to previous research, we applied unidimensional and multidimensional versions of the rmGPCM, which additionally parameterizes discrimination power of items and showed better data fit than the mPCM with equal item discrimination parameters. Researchers can use this model to evaluate the appropriateness of a selected rating scale at the stage of measure development. A task for future research may be to conduct replication studies to generalize these findings to other traits or attitudes.

[11] The ORS class in the experimental study (Chapter 4) is largely identical to the DRS class from the study about the adequacy of an 11-point rating scale (Chapter 2).

5.2 Optimal Number of Response Categories

Optimizing the rating scale is one of the methods for eliminating ICU effects (e.g., Cox, 1980). More specifically, shortening rating scale length tends to reduce the difficulties that respondents experience when determining the meaning of many fine categories, especially if only extreme categories are labeled (e.g., Krosnick, 1991; Tourangeau et al., 2000; Viswanathan et al., 2004). Previous research has not suggested any universal number of response categories because the optimal rating scale length depends strongly on the complexity of a trait or attitude of interest and sample characteristics (e.g., Dolnicar & Grün, 2009). However, a rating scale with four to six or seven response categories seems to be suitable for most research questions in the field of social and behavioral sciences (e.g., Chang, 1994; Culpepper, 2013; Lozano et al., 2008; Weng, 2004).

The goal of the experimental study (Chapter 4) was to examine whether short rating scales with four or six response categories are valid for assessing job satisfaction, compared with an 11-point rating scale. For this purpose, we recruited MTurk workers (*N* = 6,999) to participate in this study and randomly assigned them to one of three conditions in which 12 items of job satisfaction were offered with different rating scale lengths (4, 6, or 11 response categories). As mentioned above, a multidimensional rmGPCM was applied within each condition to detect patterns of ICU. Our hypothesis about shorter rating scales leading to less ICU was supported in the present study. Specifically, the proportion of respondents with ICU (non-ERS class and ERS class) decreased from 51% to 44% and 38% with 11-, 6-, and 4-point rating scales, respectively. In addition, respondents ignored less or hardly any categories when confronted with shorter rating scales. These findings can be explained by the fact that respondents can better understand short rating scales than long ones. Thus, these findings suggest that shortening the rating scale to four or six response categories is an effective strategy to eliminate a substantial extent of ICU due to the sub-optimal features of a rating scale.

However, other findings of the present study indicate that optimizing the rating scale alone is limited if ICU is caused by stable respondent characteristics. For instance, regardless of rating scale length, we found the same constellation of latent classes: the ORS class, ERS class, and non-ERS class. As such, in all conditions, respondents consistently reacted to the job satisfaction items in three ways: with ordinary responses, an overuse of extreme categories, or avoidance of extreme categories. Furthermore, the proportion of the respondents who used the ERS was nearly the same for all rating scale lengths (28% with 11- and 6-point rating scales and 26% with a 4-point rating scale). These findings accord with those of previous research, emphasizing that respondents can consistently use different rating scales in their own way (e.g., Kieruj & Moors, 2013; Moors, Kieruj, & Vermunt, 2014). Primarily, this evidence refers to the ERS.

In summary, these findings lead to the conclusion that researchers should consider both measure-dependent and person-dependent causes of ICU. With the optimization of the features of a rating scale, response-format-related ICU can be entirely eliminated; however, person-dependent effects of ICU should be controlled by suitable statistical methods (as presented in the section "Methods to Control Inappropriate Category Use"). Thus, we recommend data on job satisfaction (and other aspects of cognitive well-being) to be collected with 6-or 4-point rating scales. In particular, a 6-point rating scale may allow more reliable data to be obtained. Future studies should examine further benefits of optimizing rating scales (for instance, with regard to endpoint-labeled versus fully-labeled rating scales). We further suggest examining whether an alternative response format for data collection (e.g., item-specific response format, forced-choice response format, and ranking) is more beneficial.

5.3 Update of the Concept of Inappropriate Category Use

Inappropriate category use is a broad construct that primarily comprises RSs, shortcut strategies, social desirability, and careless responses. Response styles comprise individuals' tendencies to misuse specific response categories, for example, extreme categories (ERS), or the middle category (MRS), or the tendency to agree with items (ARS) or disagree with them (DRS) (e.g., Van Vaerenbergh & Thomas, 2013). Next, shortcut strategies are tendencies to simplifying response format, for example, by ignoring superfluous response categories when responding to items (e.g., Krosnick, 1999). Finally, respondents can respond to items in a socially desirable way (e.g., Paulhus, 1984) or in a careless way (Curran, 2016).

The use of a particular type of the ICU can be evoked by different causes. For instance, RSs can be affected by sub-optimal features of response format and person characteristics, whereas shortcut strategies are often used when a sub-optimal rating scale is offered. Socially desirable responses are expected in socially relevant contexts or when a questionnaire refers to socially sensitive topics (Zickar & Gibby, 2006), whereas careless responding is most likely evoked by contextual factors such as fatigue, distraction, or lack of interest in the topic or motivation (Curran, 2016). In the research literature, different terms such as "response style" and "response set" were used to denote different origins of ICU (for more details, see Van Herk, Poortinga, & Verhallen, 2004; Wetzel, Böhnke, & Brown, 2016). Response styles has stimuli-related causes (e.g., the wording of items, response format, time pressure or motivation), whereas response set or response tendency is person-specific (e.g., black-and-white-thinking). However, these terms are used inconsistently in the research literature. Therefore, due to multiple factors that affect respondents' responding, multiple types of ICU can simultaneously be present in the data. However, this issue has received little attention in previous research. Furthermore, most studies have examined only traditional RSs (e.g., ERS, MRS, ARS, and DRS). Few studies have considered atypical forms of RSs, such as the semi-ERS (the tendency to misuse two lower and two upper extreme categories) or the non-ERS (the tendency to avoid extreme categories). Chapters 2 and 4 of this dissertation provide supporting evidence for the complexity of ICU and the presence of atypical RSs in data.

In Chapter 4, we found the simultaneous occurrence of multiple types of ICU. First, as reported above, we identified three RS classes, regardless of rating scale length. Second, RSs were accompanied by avoided response categories, mostly when a long rating scale was offered. Thereby, the largest number of avoided categories was found in the ERS class. Third, latent classes differed in the presence of careless responding. For example, the non-ERS class showed a strong tendency to inattentive responding (max. 13%), quick responding (max. 11%), and invariant responding (max. 12%). By contrast, random responding was more present in the ERS class (max. 10%) and the ORS class (max. 8%). Fourth, the ERS class showed a higher level of social desirability in the forms of self-deceptive enhancement (11- and 6-category conditions) and impression management (11-category condition) than did the ORS class or the non-ERS class. Apparently, multiple types of ICU primarily occur when a long rating scale is used for data collection.

Both Chapters 2 and 4 provided evidence for the presence of atypical RS types in the data. In Chapter 2, apart from a latent class with traditional ERS, one of the latent classes used the semi-ERS. This form of the ERS is not new (see also Morren, Gelissen, & Vermunt, 2013), suggesting that researchers should be cautious when defining the ERS within statistical models to control RS effects. Furthermore, in Chapter 4, a latent class with the non-ERS was identified, denoted by avoidance of extreme categories. Wetzel, Carstensen, and Böhnke (2013) also reported on the non-ERS in respondents' responses to multiple trait scales assessed within the Organization for Economic Cooperation and Development's (OECD's) Programme for International Student Assessment (PISA) study (OECD, 2006). Interestingly, atypical RSs were previously identified in studies that used statistical approaches to allow for the detection of patterns of ICU a posteriori, by interpreting estimated item parameters (such as the mixture IRT approach and latent class factor approach).

Considering the findings presented above, the concept of ICU should be updated. Consequently, it is important that researchers apply different statistical methods to eliminate the effects of different types of ICU from the data and not only focus on, for instance, a few RSs, as has usually been the case in previous research. Furthermore, because both traditional and atypical RSs can be present in the data, researchers primarily need statistical methods that allow them to detect ICU in an explorative way. Controlling ICU with ad hoc methods alone does not ensure that all RSs are correctly identified.

5.4 Personal Profiles of Response-Style Users

Apart from measure-dependent factors, the use of RSs can be determined by stable respondent characteristics (e.g., social-demographic variables, big five personality traits, cognitive ability) and contextual factors. As shown in Table 1.5 (Chapter 1), previous research has provided inconsistent findings on this issue, so that no clear picture can be obtained about a personal profile of the respondents who consistently use a specific RS (for references, see Table 1.5). This inconsistency may arise for several plausible reasons, such as the inclusion of a few predictors in the analyses and the assessment of RSs based on the items measuring different traits by means of different response formats that vary across studies. In the present dissertation, we have aimed to investigate personal profiles of RS users in a systematic way.

In Chapter 4, we compared profiles of respondents showing identical RSs in three experimental conditions where the items of job satisfaction were offered with 4-, 6-, or 11-point rating scales, respectively. To describe these profiles, a large set of relevant predictors, such as socio-demographic variables, personality traits (e.g., big five personality traits, general self-efficacy), cognition-related attitudes (e.g., need for cognition, tolerance to ambiguity), cognitive ability (in the terms of the amount of working memory and the richness of vocabulary), and job-related factors (e.g., job position, organization size) were included in multinomial logistic regressions. As a result, we found two groups of statistically significant predictors: (a) *general predictors*[12] that could consistently predict the use of a particular RS, regardless of rating scale length, and (b) *response-format-specific predictors* that could explain RS only for a specific rating scale length. Thus, ERS users were generally characterized by higher general self-efficacy and perceived job autonomy, as well as higher self-deception and lower job-related stress. They differed from ORS users in job positions (a low and medium job position are more likely for ERS use) and from non-ERS users in higher neuroticism and need for cognition. Non-ERS users generally showed a lower need for cognition and experienced lower job security than did ORS users. Interesting, most socio-demographic variables, unmentioned personality traits, considered cognitive abilities, and many job-related factors appear as response-format-specific predictors, suggesting that the characteristics of respondents who use a particular RS as an adjustment strategy can be distinguished by rating scale length. In this dissertation, age and extraversion were not associated with any RS by controlling for other predictors.[13]

[12] This group of predictors also included so-called *almost general predictors* that explain the use of a particular RS in two of three conditions.

[13] Chapter 2 provides the profiles of RS users when job satisfaction is measured only with an 11-point rating scale. In accordance with the findings reported in Chapter 4, the ERS users in Chapter 2 tended to be women who worked in a low

In sum, it can be concluded that each class of RS users is heterogeneous and consists of at least two subgroups: (a) respondents who use a particular RS independent of rating scale length and (b) those who use RS to overcome difficulties caused by sub-optimal features of a rating scale. Each of these subgroups is differently characterized. The present dissertation suggests that the characteristics of the last subgroup vary across rating scales. This heterogeneity may be the major reason why previous findings concerning predicting RSs are inconclusive. Further studies should collect more evidence of the profiles of both consistent RS users and response-format-specific RS users. Furthermore, all predictors considered could explain maximally 20% of the variability in assignment to the RS classes. This finding is consistent with previous studies reporting that both socio-demographic variables and personality traits explain a small portion of RS variance (for review, see Van Vaerenbergh & Thomas, 2013). This finding emphasizes the need to extend the list of potential predictors, such as further respondent characteristics and contextual factors (e.g., motivation, knowledge about the topic, distractors). Recent studies have also highlighted the relevance of culture for RS use (Van Vaerenbergh & Thomas, 2013).

5.5 Requirements and Challenges of Mixed Item Response Theory Models

The research literature suggests diverse statistical approaches to assess ICU and to control for its effects. Basically, they can be classified into two groups: (i) *ad hoc* approaches that require researchers to make a priori assumptions about what RSs may be present in the data (e.g., confirmatory factor analysis with a latent method factor, multi-process item response theory [IRT] models or multidimensional IRT models) and (ii) *post hoc* approaches that allow them to interpret RSs from estimated item parameters resulting from the application of a certain model (e.g., mixed IRT models). In this dissertation, the mixture polytomous IRT approach was applied because of its relevant advantage of allowing for the detection of ICU patterns in an explorative way. However, the mixture polytomous IRT approach is not without its limitations. For instance, most mixed polytomous IRT models are relatively complex (especially if they include two item parameters), and therefore their application requires a large sample size (see Embretson & Reise, 2013). Subsequent studies have examined this issue primarily for mixed dichotomous IRT models (e.g., Cho, Cohen, & Kim, 2013; Dai, 2013; Finch & French, 2012; Meyer, 2010; Preinerstorfer & Forman, 2012) and the mPCM, the most parsimonious mixed polytomous IRT model (see Cho, 2014).

job position and experience higher job autonomy than ORS users (called in Chapter 2 as the DRS users). Some differences in the profile of the ERS users found in two studies can be explained by a different set of predictors considered. In the study from Chapter 2, only socio-demographic variables and job-related were included. Therefore, we found more statistically significant predictors among socio-demographic variables (e.g., low educational levels) and contextual factors (e.g., part-time work, importance of the job) than is presented in the results reported in Chapter 4.

Little knowledge about conditions under which these models work effectively discourages researchers from applying mixed polytomous IRT models. Furthermore, it is not clear how these models perform when applied to data collected in panel surveys.

Chapter 3 presents the Monte Carlo simulation study in which the performance of two mixed polytomous IRT models well-established for detecting ICU was examined under varying sample sizes (from 500 to 5,000 cases). It refers to the mixed partial credit model (mPCM; Rost, 1997) and the restricted mixed generalized partial credit model (rmGPCM; GPCM; Muraki, 1997; mGPCM; von Davier & Yamamoto, 2004). To ensure ecological validity, this simulation study used empirically derived population parameters that reflect the latent mixture of three latent classes differing by RSs and the number of avoided response categories. The primary goal was to identify the minimally required sample size for an accurate application of these two models to data collected with a long rating scale. This data situation is typical for panel surveys, and therefore the application of a mixed polytomous IRT model for detecting ICU is reasonable in this context. In addition, several information criteria (AIC, BIC, CAIC, AIC3, and SABIC) were examined regarding their effectivity to correctly identify the best-fitting class solution of each mixed IRT model. For both models, a sample size of at least 2,500 observations was required to obtain accurate parameter and standard error estimates, suggesting that these models can be straightforwardly applied in survey research. However, we warn researchers against applying these models with a sample consisting of less than 1,500 observations, and samples that include up to 2,500 observations are questionable.

Most estimation problems can arise due to an unknown latent mixture: specifically, (i) if the latent mixture is highly complex (includes more than two latent classes), (ii) if some latent classes are small-sized, and (iii) if the expected frequency of some response categories is near null (known as the sparse table problem). Further enlargement of the sample size led to the improvement of estimation accuracy and, therefore, may be effective in overcoming these estimation problems if they occur. These findings are generally consistent with previous research regarding mixed dichotomous IRT models (e.g., Preinerstorfer & Forman, 2012) and unidimensional IRT models (DeMars, 2003; He & Wheadon, 2013). Notably, our simulation study showed that the mPCM generally performed only slightly better than the rmGPCM, indicating that the model assuming varied item discrimination parametersmay not require a larger sample than when item discrimination parameters are set equal.[14] Finally, among information criterion, both the AIC3 and SABIC showed a high accuracy rate with a medium-large sample (from 1,500 and 2,500 observations, respectively). By contrast, the BIC and CAIC performed properly only with large samples (from 4,500 and 5,000 observations, respectively). The AIC showed insufficient

[14] The item discrimination parameters of the rmGPCM are restricted to be equal across the latent classes.

accuracy (≤ 86%) in all sample-size conditions. These findings have important practical implications (that will be discussed below in the "Implication Section").

A major contribution of all simulation studies is that they provide an overview of the requirements and challenges of a particular model when applied under different conditions. Such an overview can be used in the preparation stage of an empirical study. In the simulation study in this dissertation, we revealed central sample-size requirements for two mixed unidimensional polytomous IRT models. Moreover, recommendations are made for the selection of suitable information criteria for these models.[15] However, researchers should be cautious using our findings and recommendations for planning their analysis, for two reasons: (i) The latent mixture is a priori unknown and results from the model application, and (ii) model specification may affect sample size required, specifically when a researcher considers applying a more complex model, as examined in the present dissertation (e.g., the mGPCM [von Davier & Yamamoto, 2004] or the mixed nominal response model [NRM; Bock, 1972]). The first model (mGPCM) assumes that class-specific item discrimination parameters are freely estimated, and therefore this model is much more complex than the rmGPCM. The last model (mNRM) additionally includes freely estimated category discrimination parameters. Future studies may examine reasonable data conditions for applications of these complex models. Moreover, researchers may investigate the performance of mixed polytomous IRT models in the context of other features of the latent mixture (e.g., other RSs in classes, strongly distinguishing class sizes) and data situations (e.g., multidimensional scales).

5.6 Generalizability

Our findings with regard to the adequacy of an 11-point rating scale and the optimal rating scale length primarily hold for job satisfaction. It may be generalized to other aspects of cognitive well-being without any constraints. Researchers should be careful to generalize these findings to other attitudes and traits because respondents' response behavior is construct-dependent and may be affected by the content of items and complexity of the construct of interest (e.g., Cabooter, Weijters, De Beuckelaer, & Davidov, 2017; Dolnicar & Grün, 2009). Moreover, response behavior is population-dependent. For example, it is well known that students can better cope with a longer rating scale than respondents drawn from the general population (e.g., Cox, 1980; Weathers et al., 2005). The generalization is also limited to the rating

[15] The simulation study was conducted to clarify the optimal sample size for an application of mixed unidimensional polytomous IRT models in the experimental study. However, due to the multidimensionality of the job satisfaction items in the experimental study, a more complex model (the multidimensional rmGPCM) was applied within experimental conditions, as examined in the simulation study. Consequently, other information criteria than those specifically recommended in the simulation study appear to be more suitable to identify the best-fitting class solution of the multidimensional rmGPCM.

scales with endpoint labeling and positively numbered response categories. Differently designed rating scales may be more or less demanding for respondents because they provide them with cues that may be more or less helpful to accurately determine the meaning of response categories (Cabooter, Weijters, Geuens, & Vermeir, 2016; Moors et al., 2014; Tourangeau, Couper, & Conrad, 2007). Specifically, compared to fully-labeled rating scales, endpoint-labeled rating scales comprise fewer clear cues: only labels of extreme categories and the ascending numerical values of all response categories. Thus, endpoint-labeled rating scales should include fewer response categories than fully-labeled rating scales to avoid cognitive overload and any ICU (Hamby & Levine, 2016; Moors et al., 2014; Weijters, Cabooter, & Schillewaert, 2010). However, excessively long fully-labeled rating scales may be demanding (Krosnick & Fabrigar, 1997). Next, positively numbered unipolar rating scales are more demanding than are unipolar rating scales, with both positively and negatively numbered response categories because the former are generally perceived by respondents as less symmetric than the latter (Cabooter et al., 2016; Moors et al., 2014). Finally, the effect of rating scale length on ICU, which is found for even-numbered rating scales, may be strengthened by the presence of the middle category in odd-numbered rating scales (Kieruj & Moors, 2010; Moors, 2008; O'Muircheartaigh, Krosnick, & Helic, 1999; Weijters et al., 2010).

Our findings with regard to the sample-size requirements are valid primarily for both the mPCM and the rmGPCM. Both models are relatively parsimonious compared with other mixed polytomous IRT models such as the mNRM, mGPCM or further extended mixed IRT models with random-RS-factors. Researchers should be aware that the sample size of 2,500 observations may be insufficient to obtain accurate estimates of model parameters. Furthermore, our simulation study focused on a specific data situation (short test and long rating scale) and considered the latent mixture including three unequally-sized latent classes with different category use. Therefore, researchers should be cautious about using our recommendations of required sample size when they expect to obtain a more complex latent mixture (e.g., four latent classes or more, several small latent classes or several latent classes with quite similar category use). Finally, our recommendations for effective information criteria are primarily valid for mixed unidimensional IRT models.

5.7 Implications

One of the important tasks of panel surveys is to collect accurate data. This task implies practical actions that improve the psychometric quality of the measures included. On the one hand, collecting data with optimized measures is reasonable to prevent measure-dependent ICU effects (e.g., due to sub-optimal response format or unclearly formulated items). On the other hand, applying suitable statistical methods allows for the elimination of person-dependent ICU effects.

5.7.1 Optimizing the Rating Scale in Survey Research

A recent trend in panel surveys is to include short-scale measurements that are usually offered with a large number of response categories. This extension of rating scale length aims to compensate for potentially low reliability due to the small number of items (e.g., Maydeu-Olivares et al., 2009; Willits et al., 2016). However, this practice is not without its concerns. The present dissertation provides two implications that may be relevant for the inclusion or development of further measures in panel surveys. First, a long rating scale (here, 11 categories) was shown to be of limited adequacy for a valid assessment. It is open to speculation whether there are some respondents who can cope well with a long rating scale. If this is the case, it may be necessary to offer respondents a selection of rating scales which, for example, differ in the number of response categories. Second, the results suggest that a large majority of respondents use shorter rating scales consisting of six and four response categories in an adequate way. Respondents exhibit less ICU when they understand the meaning of separate response categories and can undoubtedly choose one that best reflects their judgments concerning a particular item (e.g., Krosnick, 1991). Thus, the present dissertation provides evidence that optimizing the rating scale is an effective approach to prevent measure-dependent ICU effects.

5.7.2 Dealing with Person-Dependent Inappropriate Category Use

Alternatively, several statistical methods aim to eliminate the effects of ICU (see Table 1.6 in Chapter 1). As such, it is particularly relevant to know what types of ICU may be present in the data and what are their causes. The results of this dissertation show that traditional and atypical RSs (e.g., semi-ERS), shortcut strategies (e.g., in the form of avoiding superfluous response categories), and types of careless responses can simultaneously be present in the data. Specifically, shortcut strategies (here, avoiding superfluous response categories) are primarily used to manage an unnecessarily long rating scale, whereas RSs can also be caused by stable personal dispositions, regardless of features of a rating scale. Specifically, in Chapter 4, ERS use was consistently predicted, for example, by general self-efficacy. Therefore, effects due to shortcut strategies can be prevented by optimizing the rating scale (as reported above), whereas the person-dependent effects of ICU should be eliminated using a suitable statistical approach. Thus, instead of focusing on a single or few traditional RSs, it may be more beneficial and effective to apply statistical approaches that can detect ICU in an exploratory way (e.g., mixed IRT models).

5.7.3 Applying Mixed Item Response Theory Models

The present dissertation contributes knowledge about the performance of mixed polytomous item response theory (IRT) models when applied to the data situation typical for national surveys: This result suggests that these models can effectively work with a sample, including at least 2,500 observations.

Researchers should think carefully before applying mixed IRT models to smaller samples. Applying a relatively complex mixed IRT model to a relatively small sample size may provoke estimation problems and bias model parameters. Generally, researchers have greater flexibility with a large sample. Researchers may, therefore, apply these models to evaluate the quality of the rating scale and obtain adjusted individual trait values. Furthermore, researchers may follow our recommendations concerning the selection of information criteria to accurately determine the number of latent classes if the results from different information criteria used are not consistent. Choosing a model solution with fewer classes (as the BIC often suggests) than the true number of latent classes may have an effect on person parameter estimates (e.g., Maij-de Meij, Kelderman, & van der Flier, 2008).

5.8 Conclusion and Directions for Future Research

Suggestions for future research can be proposed based on the findings of the present dissertation. Below, the most promising of these are outlined:

(1) *Replications studies on an optimal rating scale length.* This dissertation has shown that cognitive well-being, namely job satisfaction, can be accurately assessed with a shorter rating scale including six or four response categories. By contrast, a long rating scale consisting of 11 categories is of limited adequacy for a valid assessment because it produces a large amount of ICU. Low ICU was considered here as an indicator of the quality of a rating scale. Future studies should focus on other constructs such personality traits because the presence of ICU may vary depending on the construct of interest. In addition, it may be worthwhile to examine whether an optimal number of response categories found in this dissertation for the rating scale with endpoint labeling and positive numeric values remains the same for rating scales with other features (e.g., fully labeling, negatively and positively numbering, the absence of numeric labels). It is known from previous research that rating scale features can differentially affect ICU.

(2) *Explaining the use of RSs.* Respondents can substantially differ in the reasons they use a particular RS. To explain these differences, the measure itself (e.g., item wording, features of response format, scale length), contextual factors (e.g., motivation to participate, knowledge about the topic, mood, fatigue, distractors, and time pressure), and dispositional variables or further stable personal characteristics (e.g., personality traits, cognition-related attitudes, attitude towards accurate responding, cognitive ability, and socio-demographic variables) should be considered. In addition, culture may be added to the list of relevant factors that determine consistent RS use.

(3) *Self-selected rating scale.* Respondents can differ in their preferences of the rating scale they consider optimal. Specifically, more educated respondents (e.g., students) can successfully cope with more demanding rating scales (e.g., endpoint-labeled rating scale, long rating scales) than can less educated

respondents. A further worthwhile suggestion for future research is to examine whether the amount of ICU would diminish if respondents were to answer questions with the self-selected rating scale. A self-selected rating scale may optimally represent the respondent's subjective categories and facilitate a successful transfer of his or her answers into the rating scale. Thus, it is likely that answering questions with an individually optimal rating scale would remedy measure-dependent ICU.

(4) *Alternative response format.* To prevent ICU, researchers can collect data with an alternative response format (e.g., item-specific response format, funnel response format, drag and drop response format, forced-choice response format, ranking, or magnitude estimation scale). In contrast to rating scales, the alternative response scales generally trigger less complex response processes, are more user-friendly (e.g., due to their direct design, less cognitive burden by responding, and quicker response time), evoke less ICU (Böckenholt, 2017; Dolnicar & Grün, 2009; Harzing, Baldueza, Barner-Rasmussen, Barzantny, Canabal, Davila, ... & Liang, 2009; Khorramdel & von Davier, 2014; Saris, Revilla, Krosnick, & Shaeffer, 2010), and exceed rating scales in terms of psychometric data quality (e.g., Brown & Maydeu-Olivares, 2011; Churchill & Peter, 1984; Harzing et al., 2009; Koskey, Sondergeld, Beltyukova, & Fox, 2013; Saris et al., 2010).

(5) *Methodological innovations for measuring ICU.* Multiple types of ICU can be present in the data. Moreover, RSs can occur both in traditional or atypical form, especially in the presence of a long rating scale. To control their effects, a statistical approach is needed that can both detect patterns of ICU and qualify their effects. It may be worthwhile to proceed using the following two steps: (1) patterns of ICU are detected using an exploratory approach such as mixed polytomous IRT models, and (2) information about identified patterns is used for specifying RS factors within one of the ad hoc models to quantify these RSs and control their effects at the individual level (e.g., multidimensional IRT models with a trait factor and multiple RS factors). Recently, Huang (2016) proposed the mixed GPCM model with a random-effect factor to allow for simultaneously unmixing a sample into latent classes with different RSs and quantify the respondent's intensity of the ERS. However, this model does not include multiple random-effect factors to quantify multiple RSs.

An important issue in survey research is the assessment of valid data. The present dissertation has shown that a long rating scale can be a serious source of ICU effects and has provided evidence that shorter rating scales are more adequate for obtaining valid data. Furthermore, this dissertation has demonstrated the benefits of an optimization of rating scale length to eliminate the measure-related effects of ICU. In addition, the present dissertation has contributed to clarifying personal profiles of RS users. Finally, this dissertation has clarified the sample-size requirements and challenges of applying mixed polytomous IRT models to demanding data.

5.9 References

Baumgartner, H., & Steenkamp, J.-B. E. M. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research, 38*, 143-156. doi: 10.1509/jmkr.38.2.143.18840

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*(1), 29-51. doi: 10.1007/BF02291411

Böckenholt, U. (2017). Measuring response styles in Likert items. *Psychological Methods*, *22*(1), 69-83. doi: 10.1037/met0000106

Brown, A., & Maydeu-Olivares, A. (2011). Item response modeling of forced-choice questionnaires. *Educational and Psychological Measurement*, *71*(3), 460-502. doi: 10.1177/0013164410375112

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. doi:10.1007/s11135-016-0411-5

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. doi: 10.1016/j.jbusres.2015.10.138

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*, 205-215. doi: 10.1177/014662169401800302

Cho, S. J., Cohen, A. S., & Kim, S. H. (2013). Markov chain Monte Carlo estimation of a mixture item response theory model. *Journal of Statistical Computation and Simulation*, *83*, 278-306. doi: 10.1080/00949655.2011.603090

Cho, Y. (2014). The mixture distribution polytomous rasch model used to account for response styles on rating scales: A simulation study of parameter recovery and classification accuracy. *Dissertation Abstracts International*, *75*. http://hdl.handle.net/1903/14511

Churchill, G. A., Jr., & Peter, J. P. (1984). Research design effects on the reliability of rating scales: A meta-analysis. *Journal of Marketing Research, 21*, 360-375.

Cox, E. P. (1980). The optimal number of response alternatives for a scale: A review. *Journal of Marketing Research, 17*(4), 407-422. doi: 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*, 201-225. doi: 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. doi: 10.1016/j.jesp.2015.07.006

Dai, Y. (2013). A mixture Rasch model with a covariate: A simulation study via Bayesian Markov chain Monte Carlo estimation. *Applied Psychological Measurement*, *37*, 375-396. doi: 10.1177/0146621612475076

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*, 104–115. doi: 10.1509/jmkr.45.1.104

DeMars, C. E. (2003). Sample size and the recovery of nominal response model item parameters. *Applied Psychological Measurement*, *27*, 275-288. doi: 10.1177/0146621603027004003

Diener, E., Inglehart, R., & Tay, L. (2013). Theory and validity of life satisfaction scales. *Social Indicators Research*, *112*(3), 497-527.

Dolnicar, S., & Grün, B. (2009). Does one size fit all? The suitability of answer formats for different constructs measured. *Australasian Marketing Journal*, *17*(1), 58-64.

Embretson, S. E., & Reise, S. P. (2013). *Item response theory for psychologists*. Hillsdale, New Jersey: Erlbaum.

Finch, W. H., & French, B. F. (2012). Parameter estimation with mixture item response theory models: A Monte Carlo comparison of maximum likelihood and Bayesian methods. *Journal of Modern Applied Statistical Methods*, *11*, 167-178. doi: 10.22237/jmasm/1335845580

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. doi: 10.1080/1091367X.2013.807265

Hamby, T., & Levine, D. S. (2016). Response-scale formats and psychological distances between categories. *Applied Psychological Measurement*, *40*(1), 73-75. doi: 10.1177/0146621615597961

Harzing, A. W., Baldueza, J., Barner-Rasmussen, W., Barzantny, C., Canabal, A., Davila, A., ... & Liang, Y. K. (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. doi: 10.1016/j.ibusrev.2009.03.001

He, Q., Wheadon, C. (2013). The effect of sample size on item parameter estimation for the partial credit model. *International Journal of Quantitative Research in Education*, *1*, 297-315. doi: 10.1504/IJQRE.2013.057692

Huang, H. Y. (2016). Mixture random-effect IRT models for controlling extreme response style on rating scales. *Frontiers in Psychology, 7*, 1706. doi:10.3389/fpsyg.2016.01706

Jin, K. Y., & Wang, W. C. (2014). Generalized IRT models for extreme response style. *Educational and Psychological Measurement, 74*, 116-138. doi: 10.1177/0013164413498876

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. doi: 10.1186/1477-7525-10-80

Khorramdel, L., & von Davier, M. (2014). Measuring Response Styles Across the Big Five: A Multiscale Extension of an Approach Using Multinomial Processing Trees. *Multivariate Behavioral Research, 49*(2), 161-177.

Kieruj, N. D., & Moors, G. (2013). Response style behavior: question format dependent or personal style?. *Quality & Quantity*, *47*(1), 193-211. doi: 10.1007/s11135-011-9511-4

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. doi: 10.1093/ijpor/edq001

Koskey, K. L., Sondergeld, T. A., Beltyukova, S. A., & Fox, C. M. (2013). An experimental study using Rasch analysis to compare absolute magnitude estimation and categorical rating scaling as applied in survey research. *Journal of Applied Measurement*, *14*(3), 262-281.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology, 5*(3), 213-236. doi: 10.1002/acp.2350050305

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology, 50*, 537-567. doi: 10.1146/annurev.psych.50.1.537

Krosnick, J. A. & Fabrigar, L. R. (1997). Designing rating scales for effective measurement in surveys. In L. Lyberg, P. Biemer, M. Collins, L. Decker, E. de Leeuw, C. Dippo, et al. (Eds.), *Survey Measurement and Process Quality* (pp. 141–164). New York: John Wiley & Sons, Inc.

Krosnick, J.A., & Presser, S. (2010). Question and questionnaire design. In P.V. Marsden, J.D. Wright (Eds.), *Handbook of survey research* (pp. 263-314). Bingley, UK: Emerald Group Publishing.

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. doi: 10.1177/0734282914522200

Lozano, L. M., García-Cueto, E., and Muñiz, J. (2008). Effect of the number of response categories on the reliability and validity of rating scales. *Methodology* 4, 73-79.

Maij-de Meij, A. M., Kelderman, H., & van der Flier, H. (2008). Fitting a mixture item response theory model to personality questionnaire data: Characterizing latent classes and investigating possibilities for improving prediction. *Applied Psychological Measurement, 32*, 611-631. doi:10.1177/0146621607312613

Maydeu-Olivares, A., Kramp, U., García-Forero, C., Gallardo-Pujol, D., and 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. doi: 10.3758/BRM.41.2.295

McCredie, M. N., & Morey, L. C. (2018). Who are the Turkers? A characterization of MTurk workers using the personalityassessment inventory. *Assessment,* 1–8. doi: 10.1177/1073191118760709

Meyer, J. P. (2010). A mixture Rasch model with item response time components. *Applied Psychological Measurement, 34*, 512–538. doi: 10.1177/0146621609355451

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*, 81–97. doi: 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. doi: 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. doi: 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.

Muraki, E. (1997). A generalized partial credit model. In W. van der Linden & R. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 153-164). New York: Springer.

Naemi, B. D., Beal, D. J., & Payne, S. C. (2009). Personality predictors of extreme response style. *Journal of Personality, 77*(1), 261-286. doi: 10.1111/j.1467-6494.2008.00545.x

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

OECD (2006). *PISA 2006 assessment framework*. Paris, France: OECD Publications.

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology*, *46*(3), 598-609.

Preinerstorfer, D., & Formann, A. K. (2012). Parameter recovery and model selection in mixed Rasch models. *British Journal of Mathematical and Statistical Psychology*, *65*, 251-262. doi: 10.1111/j.2044-8317.2011.02020.x

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*, 1-15. doi:10.1016/S0001-6918(99)00050-5

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. doi: 10.1177/0049124113509605

Rost, J. (1997). Logistic mixture models. In W. van der Linden & R. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 449-463). New York: Springer.

Saris, W., Revilla, M., Krosnick, J. A., & Shaeffer, E. M. (2010, May). Comparing questions with agree/disagree response options to questions with item-specific response options. In *Survey Research Methods* (Vol. 4, No. 1, pp. 61-79). doi: 10.18148/srm/2010.v4i1.2682

Shaftel, J., Nash, B. L., & Gillmor, S. C. (2012, April). 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).

Summerfield, M., Bevitt, A., Freidin, S., Hahn, M., La, N., Macalalad, N., ... & Wooden, M. (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. doi: 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). Springer, Cham.

Tourangeau, R., Couper, M.P., & Conrad, F. (2007). Colors, labels, and interpretive heuristics for response scales. *Public Opinion Quarterly, 71*, 91–112. doi: 10.1093/poq/nfl046

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response.* New York, NY US: Cambridge University Press.

Van Herk, H., Poortinga, Y. H., & Verhallen, T. M. M. (2004). Response styles in rating scales: Evidence of method bias in data from six EU countries. *Journal of Cross-Cultural Psychology, 35*(3), 346–360. doi: 10.1177/0022022104264126

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. doi: 10.1093/ijpor/eds021

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. doi: 10.1016/s0148-2963(01)00296-x

von Davier, M., & Yamamoto, K. (2004). Partially observed mixtures of IRT models: An extension of the generalized partial-credit model. *Applied Psychological Measurement, 28*, 389-406. doi: 10.1177/0146621604268734

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*, 1516−1524. doi: 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. doi: 10.1016/j.ijresmar.2010.02.004

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*, 956-972. doi: 10.1177/0013164404268674

Wetzel, E., Böhnke, J. R., & Brown, A. (2016). Response biases. In F. T. L. Leong, D. Bartram, F. Cheung, K. F. Geisinger, & D. Iliescu (Eds.), *The ITC international handbook of testing and assessment* (pp. 349–363). Oxford, England: Oxford University Press.

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. doi: 10.1016/j.jrp.2012.10.010

Willits, F. K., Theodori, G. L., & Luloff, A. E. (2016). Another look at Likert scales. *Journal of Rural Social Sciences, 31*(3), 126-139.

Zickar, M., & Gibby, R. E. (2006). A history of faking and socially desirable responding on personality test. In R. L. Griffith & M. H. Petterson (Eds.), *A closer examination of applicants faking behavior* (pp.21-42). Charlotte, NC: IAP-Information Age Publishing.

6 LIST OF TABLES

7 List of Figures

8 APPENDIX IN GERMAN LANGUAGE

8.1 Zusammenfassung

Arbeitszufriedenheit ist ein Aspekt des kognitiven Wohlbefindens und gilt in der Sozialforschung als Standardindikator für Lebensqualität. Die Erfassung der Arbeitszufriedenheit mit einem reliablen und validen Messinstrument ist eine Voraussetzung dafür, dass präzise Analyseergebnisse erzielt und valide Schlussfolgerungen gezogen werden können. Allerdings kann ein unangemessenes Antwortformat das Antwortverhalten von Befragten beeinträchtigen. Es besteht die Vermutung, dass eine 11-stufige Ratingskala, die in nationalen Panel-Befragungen zur Erfassung des kognitiven Wohlbefindens standardmäßig eingesetzt wird, eine Fehlerquelle sein könnte. Diese Ratingskala enthält viele feine Antwortkategorien, die zur Überforderung der Befragten führen und unangemessenes Antwortverhalten (z. B. Antwortstile, fahrlässiges Antwortverhalten, Auslassen unklarer und überflüssiger Antwortkategorien) auslösen können. Dies kann die Qualität der Daten aus Panel-Befragungen vermindern. Im Hinblick darauf sind die Ziele der vorliegenden Dissertation wie folgt festgelegt: (1) Die Tauglichkeit einer 11-stufigen Ratingskala für eine valide Erfassung der Arbeitszufriedenheit wird geprüft. (2) Es wird untersucht, wie die Mischverteilungs-IRT-Modelle für polytome Items funktionieren, unter der Datenbedingung, die für die Erfassung der Arbeitszufriedenheit in Panel-Befragungen typisch ist (wenige Items und viele Antwortkategorien). Die Mischverteilungs-IRT-Modelle haben sich für die Identifizierung von Mustern unangemessenen Antwortverhaltens etabliert. Die bisherige Forschung liefert jedoch nur relativ lückenhafte Kenntnisse über optimale Bedingungen für eine effektive Anwendung dieser Modelle. (3) Das nächste Ziel ist es, eine optimale Anzahl von Antwortkategorien für die Erfassung der Arbeitszufriedenheit zu identifizieren. Darüber hinaus befasst sich diese Dissertation mit einer weiteren Forschungsfrage: (4) Wie lassen sich Befragte, die einen bestimmten Antwortstil präferieren, anhand von Persönlichkeitseigenschaften, kognitiven Fähigkeiten, soziodemografischen Variablen und kontextuellen (hier, arbeitsbezogenen) Faktoren beschreiben? Es ist wichtig ihre Personenprofile klarzustellen, da Antwortstile über die Messinstrumente, die verschiedene Traits auch mittels unterschiedlicher Ratingskalen erfassen, konsistent auftreten können und somit auf eine dispositionsartige Ursache hindeuten.

Als Erstes wurde die Angemessenheit einer 11-stufigen Ratingskala anhand der Daten über fünf Aspekte der Arbeitszufriedenheit aus der Household, Income and Labour Dynamics in Australia (HILDA) Panel-Befragung (Welle von 2001, *n* = 7036) untersucht. Muster unangemessenen Antwortverhaltens, das ein Indikator für eingeschränkte Tauglichkeit einer Ratingskala ist, wurden mittels Mischverteilung-IRT-Modelle identifiziert. Mit diesen Modellen ließen sich drei latente Klassen mit

unterschiedlichem Antwortverhalten ermitteln. Keine der gefundenen Klassen nutzte alle elf Antwortkategorien. Die Befragten ignorierten viele Antwortkategorien. Darüber hinaus wurde eine hohe Rate an unangemessenem Antwortverhalten gefunden. Die meisten Befragten (60%) bevorzugten entweder die beiden extremsten Antwortkategorien (die latente Klasse mit einem extremen Antwortstil) oder eine Kombination aus den zwei untersten und zwei höchsten Antwortkategorien (die latente Klasse mit dem sogenannten semi-extremen Antwortstil). Das Antwortverhalten der restlichen Befragten war angemessener (die latente Klasse mit dem sogenannten differenzierten Antwortverhalten). Diese Ergebnisse deuten auf die eingeschränkte Tauglichkeit einer 11-stufigen Ratingskala zur Erfassung der Arbeitszufriedenheit hin und stellen ihre Eignung, feine interindividuelle Unterschiede in Levels der Arbeitszufriedenheit unverzerrt abbilden zu können, infrage. Die Befragten scheinen mit überflüssigen Antwortkategorien überfordert zu sein und nutzen unangemessenes Antwortverhalten, wenn sie Schwierigkeiten wegen unklarer Bedeutung einiger Antwortkategorien haben. Die Ergebnisse der Studie lassen vermuten, dass eine Ratingskala mit weniger Antwortkategorien besser ist.

Das zweite Ziel dieser Dissertation ist es, die Performanz der Mischverteilungs-IRT-Modelle für polytome Items auf Basis einer Monte-Carlo-Simulationsstudie zu untersuchen. Der Fokus lag auf zwei Modellen, die gut zur Identifizierung von Mustern unangemessenen Antwortverhaltens etabliert sind: das Mischverteilung-Partial-Credit-Modell (Rost, 1997) und das restringierte Mischverteilungs-Generalisierte-Partial-Credit-Modell (rmGPCM; GPCM; Muraki, 1997; mGPCM; von Davier & Yamamoto, 2004). Dieses Modell ist komplexer und nimmt frei zu schätzende Diskriminationsparameter an. Die Diskriminationsparameter sind jedoch in latenten Klassen invariant. Konkret war das Ziel der Simulationsstudie herauszufinden, wie groß eine optimale Stichprobengröße sein sollte, um eine ordnungsgemäße Anwendung dieser Modelle zu gewährleisten, um zwar unter der Datenbedingung, die für Panel-Befragung typisch ist. Eine weitere Forschungsfrage befasste sich damit, welche Informationskriterien (AIC, BIC, CAIC, AIC3 und SABIC) für die Identifizierung der besten Klassenlösung angemessen sind. Die Analysen zeigten, dass eine minimale Stichprobe mindestens 2500 Fälle enthalten sollte. Eine weitere Vergrößerung der Stichprobe führte zu einer besseren Schätzgenauigkeit von Parametern und Standardfehlern. Hauptsächlich zeigte die Simulationsstudie, dass das sparsame Mischverteilungs-Partial-Credit-Modell nur geringfügig besser funktionierte als das komplexere Mischverteilungs-Generalisierte-Partial-Credit-Modell. Beide Modelle hatten Schätzprobleme wegen niedriger Kategorienhäufigkeiten und produzierten in solchem Fall verzerrte Schätzwerte. Für die empfohlene Stichprobengröße erwiesen sich das AIC3 und das SABIC als am besten geeignet. Für eine große Stichprobe (ab 4500 Fälle) waren das BIC und das CAIC effektiv. Dahingegen zeigte das AIC eine unzureichende Genauigkeitsrate für alle untersuchten Stichprobengrößen.

Das dritte Ziel der vorliegenden Dissertation ist es, eine optimale Anzahl der Antwortkategorien in einer Ratingskala für eine valide Erfassung der Arbeitszufriedenheit zu identifizieren. Diese

Fragestellung wurde mittels eines experimentellen Designs mit Randomisierung untersucht (N = 6999 Arbeitnehmer aus den USA). Konkret wurden zwei Ratingskalen mit weniger Antwortkategorien (vier und sechs Antwortkategorien) mit der 11-stufigen Ratingskala verglichen. Als Kriterien galten das Auftreten unangemessenen Antwortverhaltens und die Reliabilität der Skala zur Messung der Arbeitszufriedenheit. In den drei experimentellen Bedingungen wurden die Analysen mittels eines multidimensionalen Mischverteilungs-IRT-Modells durchgeführt. Es ist zu erwähnen, dass in der Planungsphase dieser Studie die Ergebnisse aus der Simulationsstudie verwendet wurden (z. B. um die minimal erforderliche Stichprobegröße in einer experimentellen Bedingung festzulegen). Mit der Verringerung der Anzahl der Antwortkategorien in der Ratingskala reduzierten sich insgesamt der Anteil an Befragten mit unangemessenem Antwortverhalten sowie die Anzahl der ausgelassenen Antwortkategorien. Es deutet sich an, dass eine Ratingskala mit weniger Antwortkategorien im Vergleich zu einer 11-stufigen Ratingskala einen geringeren Messfehler erzeugt. Dieses Ergebnis bestätigt die Vermutung, dass einige Befragte Antwortstile nur deshalb einsetzen, um mit einer unangepassten Ratingskala zurechtzukommen. Ein weiteres interessantes Ergebnis ist, dass die gleichen Antwortstile für alle drei Ratingskalen auftraten. Es liegt nahe, dass eine Optimierung der Ratingskala nur teilweise unangemessenes Antwortverhalten verhindern kann. Anscheinend nutzen manche Befragten aufgrund von substantiellen Dispositionen bestimmte Antwortstile.

Schließlich wurden die Personenprofile von Befragten, die einen bestimmten Antwortstil nutzten (z. B. den extremen Antwortstil), beschrieben. Diese Forschungsfrage wurde mittels zweier Datensätze untersucht. Der erste Datensatz enthielt ein kleines Set von potenziellen Prädiktoren, die in der HILDA Panel-Befragung vorhanden sind. Diese sind überwiegend soziodemografische Variablen und arbeitsbezogene Faktoren. Der zweite Datensatz stammt aus der experimentellen Studie, in die vielfältige relevante Skalen und Variablen absichtlich aufgenommen wurden, wie zum Beispiel Persönlichkeitseigenschaften, kognitive Fähigkeiten, soziodemografische Variablen und arbeitsbezogene Faktoren. Jeder Datensatz enthielt eine Kriteriums-Variable, die die identifizierte Klassenzugehörigkeit der Befragten in der jeweiligen Studie abbildete. Die Analysen wurden mittels multinomialer logistischer Regressionen durchgeführt. Die Ergebnisse auf Basis der HILDA-Daten lieferten die für die 11-stufige Ratingskala spezifischen Personenprofile. Im Gegensatz dazu ließen sich auf Basis der experimentellen Daten zwei Typen von Prädiktoren identifizieren: (i) *universelle Prädiktoren*, die die Nutzung eines Antwortstils unabhängig von der Ratingskala erklären, und (ii) *antwortformat-spezifische Prädiktoren*, die das Auftreten eines Antwortstils für eine Ratingskala nur mit einer bestimmten Anzahl der Antwortkategorien erklären. Beispielweise zeigten sich eine hohe allgemeine Selbstwirksamkeit und selbst wahrgenommene Autonomie am Arbeitsplatz als universelle Prädiktoren für die Nutzung des extremen Antwortstils; niedriges Kognitionsbedürfnis wurde als universeller Prädiktor für den non-extremen Antwortstil gefunden. Der non-extreme Antwortstil bezeichnet die Tendenz zur Vermeidung der extremen Antwortkategorien. Somit stellen diese Ergebnisse einige dispositionsartige Ursachen für die Antwortstile klar. Als antwortformat-spezifische Prädiktoren wurden soziodemografische Variablen, kognitive Fähigkeiten und einige arbeitsbezogene Faktoren identifiziert. Die Ergebnisse deuten darauf hin, dass die Charakteristika der Befragten, die einen bestimmten Antwortstil nutzen, variieren können, je nachdem mit welcher Ratingskala die Daten erhoben werden. Hauptsächlich beschreiben die antwortformat-spezifischen Prädiktoren diejenige Subgruppe von Befragten, die auf eine unangepasste Ratingskala mit Antwortstilen als Anpassungsstrategie reagieren.

Zusammenfassend lässt sich festhalten, dass eine 11-stufige Ratingskala gravierende Mängel aufweist: zum Beispiel einen hohen Anteil an Befragten, die Antwortstile nutzen, sowie viele ausgelassene Antwortkategorien. Somit erwies sie sich als ungeeignet für eine valide Erfassung der Arbeitszufriedenheit. Für diesen Zweck sind die Ratingskalen mit weniger Antwortkategorien (4- und 6-stufige Ratingskalen) besser geeignet: weniger Befragten reagieren auf diese Ratingskalen in unangemessener Form. Außerdem beachten die mit diesen Ratingskalen Befragten (fast) alle Antwortkategorien eines Items. Folglich erwies sich eine Optimierung der Ratingskala als eine effektive Methode, um messinstrumentbedingtes unangemessenes Antwortverhalten zu eliminieren. Dennoch können dieselben Antwortstile unabhängig von der Ratingskala in den Daten existieren. Die Ergebnisse deuten darauf hin, dass stabile Dispositionen eine weitere relevante Ursache für das Auftreten der Antwortstile sind. Einige relevante Personencharakteristiken konnten in dieser Dissertation identifiziert werden (*universelle Prädiktoren*). Zum Beispiel lässt sich eine konsistente Nutzung des extremen Antwortstils durch eine hohe allgemeine Selbstwirksamkeit und selbst wahrgenommene Autonomie am Arbeitsplatz erklären. Diese Dissertation zeigte, dass die Antwortstile, wenn sie durch stabile Dispositionen bedingt sind, trotz einer Optimierung der Ratingskala bestehen bleiben. Mittels geeigneter statistischer Modelle können jedoch ihre Effekte kontrolliert und bereinigte Trait-Werte erhalten werden. Einer der etablierten Ansätze dafür ist der Mischverteilung-IRT-Ansatz. Die vorliegende Dissertation diskutiert die Anforderungen und Herausforderungen bei der Anwendung einiger Mischverteilung-IRT-Modelle und leitet Empfehlungen auf Basis zentraler Befunde ab.

Literatur

Muraki, E. (1997). A generalized partial credit model. In W. van der Linden & R. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 153-164). New York: Springer.

Rost, J. (1997). Logistic mixture models. In W. van der Linden & R. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 449-463). New York: Springer.

von Davier, M., & Yamamoto, K. (2004). Partially observed mixtures of IRT models: An extension of the generalized partial-credit model. *Applied Psychological Measurement, 28*, 389-406. doi: 10.1177/0146621604268734

8.2 Erklärung

Hiermit versichere ich, dass ich die vorgelegte Arbeit selbständig verfasst habe. Andere als die angegebenen Hilfsmittel habe ich nicht verwendet. Die Arbeit ist in keinem früheren Promotionsverfahren angenommen oder abgelehnt worden.

Berlin, April 2019 Tanja Kutscher