Investigating Response Styles and Item Homogeneity Using Item Response Models

Inaugural-Dissertation

in der Fakultät Humanwissenschaften

der Otto-Friedrich-Universität Bamberg

vorgelegt von

Eunike Wetzel

aus Mainz

Bamberg, den 26.03.2013

Tag der mündlichen Prüfung: 27.05.2013

Dekan	Prof. Dr. Stefan Lautenbacher
Erstgutachter	Prof. Dr. Claus H. Carstensen
Zweitgutachter	Prof. Dr. Michael Hock

Mein Dank gilt…

… Claus H. Carstensen für die Betreuung meiner Dissertation.

… Jan R. Böhnke für so viele Dinge, dass ich sie hier unmöglich alle aufzählen kann. Jan, du bist meine erste Anlaufstelle für Methodenfragen und Feedback zu Forschungsideen und Papers, mein Lieblings-Kooperationspartner und ein sehr guter Freund.

… meiner Familie Klaus, Ulrike, Susanne und Debora Wetzel für eure Liebe und Unterstützung.

… Matthias Ziegler für das Studienprojekt zu den NEOs mit dem für mich im Bereich der Persönlichkeitsdiagnostik alles angefangen hat. Danke, dass du mich auf das Thema response styles gebracht hast sowie für das Feedback zu mehreren Papers und der Synopse.

… Tim Croudace und Anna Brown für die Summer School 2010 in Cambridge, die meine Faszination für die IRT und damit das Thema dieser Dissertation maßgeblich geprägt hat.

… Michael Hock für die Übernahme des Zweitgutachtens.

… Benedikt Hell für die gute Zusammenarbeit im Genderfairness-Projekt.

… Cordula Artelt für die netten Gespräche im Mentoring.

… Fritz Ostendorf für die Daten der NEO-PI-R Normierungsstichprobe.

… meinen studentischen Hilfskräften über die Jahre: Amanda, Bogomil, Sophia, Sabrina und Benjamin. Danke, dass ihr mir die Arbeit wesentlich erleichtert haben.

Abstract

Measurement invariance is a pre-requisite for drawing accurate and valid inferences con cerning individuals' trait levels from questionnaire data. However, several factors exist that can influence a person's item responses in addition to his or her latent trait level and in consequence violate measurement invariance. The research in this dissertation was aimed at investigating three of these factors: 1) individual differences in response styles, 2) the measurement invariance of items between subgroups of respondents, and 3) the measurement invariance of items across assessment periods.

To investigate the first factor, two alternative approaches to modeling response styles were applied: the categorical approach, which posits that response styles can be understood as categorical variables, and the dimensional approach, which posits that response styles are continuous variables. In the framework of the categorical approach, mixed Rasch analyses of data from the German NEO-PI-R showed that respondents differed systematically in their response scale use: some preferred extreme categories (extreme response style) while others preferred moderate categories (non-extreme response style). In the framework of the dimensional approach, multidimensional item response models were applied to model response styles and traits simultaneously. These showed that response styles (especially extreme response style) can explain variance in item responses that is incremental to the variance explained by the traits. Thus, individual differences in response styles have an influence on item responses. This carries important implications for comparisons between individuals. Trait scores based on summing item responses should not be used to conduct trait comparisons since they can be biased when individuals differ in their response style. In contrast, both mixed Rasch models and multidimensional models can provide trait estimates that are corrected for response style effects since they allow separating response style variance from trait variance.

The second factor, the measurement invariance of items between subgroups of respondents, was investigated with respect to differential item functioning for gender in the German NEO-PI-R. Several NEO-PI-R facets especially on neuroticism (anxiety, angry hostility), agreeableness (modesty), and conscientiousness (achievement striving, deliberation) contained items showing differential item functioning for gender, indicating that these items were not measurement invariant for men and women. Differential item functioning for gender was also analyzed separately for response style groups based on the latent classes derived from mixed Rasch models. Overall, findings were consistent between the complete sample and the two response style groups (non-extreme response style and extreme response style), though some differences in the classification of differential item functioning as negligible, slight to moderate, or moderate to large as well as in the magnitude of differential item functioning occurred.

The third factor, the measurement invariance of items across assessment periods, was investigated for link items from the reading and science domains in the Programme for International Student Assessment (PISA). To this effect, data from the German PISA 2000 sample and data from a German sample that was tested in addition to the PISA 2009 sample was analyzed. Measurement invariance was violated for both item sets. For reading, this pertains to the link between PISA 2000 and PISA 2009 whereas for science, this pertains to the link between PISA 2006 and PISA 2009. Some items showed large differences in item difficulty between assessments which may in part be attributed to changes in item wording and position effects. Analyses of the link error suggest that removing items with large differences in item difficulty from the link, increasing sample sizes for the link, and maintaining item positions across assessments may reduce the link error and thus contribute to stable trends.

In sum, it was shown that individual differences in response styles, the lack of measurement invariance of items between subgroups of respondents, and the lack of measurement

invariance of items across assessment periods can impair the measurement of the intended traits and in consequence render trait inferences and comparisons between individuals or groups invalid. Thus, measures should be taken to reduce the impact of factors that interfere with measurement invariance. These measures can be aimed at test construction where, for example, the item or response format can be adjusted to elicit response styles to a lesser degree and items can be selected that have invariance properties across subgroups of participants and across assessment periods.

Contents

1. Synopsis

1.1. Introduction

One of the goals in psychological assessment focused at individuals is to draw inferences about the trait levels[1] of persons. These inferred trait levels can be used in a variety of contexts, for example, to compare the aptitude of several applicants for a job, to screen for disorders, or to conduct comparisons across age, gender, or cultural subgroups. This dissertation focusses on the context of the assessment of personality and abilities where inferences about trait levels are based on the responses of individuals to items in personality questionnaires or cognitive tests. To be able to draw accurate and valid inferences from questionnaire data, *measurement invariance* has to exist. In the broadest sense, measurement invariance means that the same measurement model holds for all respondents; i.e., relationships between items and latent traits are invariant across groups (Borsboom, Romeijn, & Wicherts, 2008; Mellenbergh, 1989; Meredith, 1993; Widaman & Reise, 1997). Measurement invariance is especially critical when the intention is to compare different groups (e.g., men and women, different age groups) regarding the trait of interest. Measurement invariance encompasses a variety of aspects that pertain to properties of the respondents and to properties of the items. Measurement invariance for example requires that 1) respondents are homogenous regarding the trait being assessed, 2) respondents use the rating scale in the same manner, and that the items are measurement invariant for 3) different subgroups of respondents and – in the case of trend analyses – 4) across assessment periods. Thus, measurement invariance can be investigated from different perspectives, either focusing on respondents or focusing on items.

[1] Throughout this dissertation, the term "trait level" refers to a person's true latent trait level, the term "trait score" refers to sum scores derived from questionnaire data, and "trait estimates" refers to estimates of latent trait levels derived from analyses of questionnaire data based on item response theory such as weighted likelihood estimates.

The research in this dissertation is aimed at investigating three factors that influence item responses in addition to an individual's trait level and in consequence threaten measurement invariance. The focus lies on *response styles* which are individual differences in response scale use that influence a person's response to an item in addition to his or her latent trait level (Wetzel, Carstensen, & Böhnke, 2013). Response styles can violate the first two requirements of measurement invariance stated above. Individual differences in response styles and the consistency of these response styles across traits are investigated using data from the German NEO-PI-R (Ostendorf & Angleitner, 2004) as well as data from the student questionnaire used in the Programme for International Student Assessment (PISA) 2006 assessment (OECD, 2006). For the NEO-PI-R data, the measurement invariance of items between subgroups of respondents is also addressed (requirement 3). Lastly, one paper examines the measurement invariance of items across assessment periods (requirement 4) for reading and science items from PISA 2000, PISA 2006, and PISA 2009. In sum, the three factors pertaining to measurement invariance investigated here are 1) individual differences in response styles, 2) the measurement invariance of items between subgroups of respondents, and 3) the measurement invariance of items across assessment periods.

The outline of this synopsis is as follows: First, I will explain the three factors in more detail and summarize previous research investigating them. Second, the role of each of the five papers this dissertation is based on in investigating the factors will be described and the main results of these papers will be summarized. The third part provides a discussion of the main results, links the findings of this dissertation to previous research in this area, and introduces possible future research questions related to the research reported here. Finally, the fourth section ends with concluding remarks. Since the focus of this dissertation is on response styles, throughout the synopsis most room will be given to the explanation and discussion of response styles.

1.2. Three factors affecting measurement invariance

The general assumption underlying inferences and comparisons between individuals or groups based on trait scores is that these trait scores accurately represent the persons' latent trait levels. However, this is only the case when an individual's latent trait level is the sole factor influencing his or her responses to the items. Other factors exist that may influence a person's responses to a certain extent. The aim of this dissertation is to investigate three of these factors: 1) individual differences in response styles, 2) the measurement in-variance of items between subgroups of respondents, and 3) the measurement invariance of items across assessment periods. The first operates on the side of the respondents while the latter two are properties of the items, though interactions between persons and items can influence all three factors.

The first factor that can affect measurement invariance is individual differences in response styles. Response styles are a response bias that is characterized by "systematic individual differences in response scale use that are independent of item content and the respondent's trait level" (Wetzel et al., 2013, p. 178; see also Paulhus, 1991). Examples of response styles are extreme response style (ERS), a preference for extreme response categories, non-extreme response style (NERS), a preference for non-extreme (i.e., moderate) response categories, acquiescence response style (ARS), a tendency to agree to statements, disacquiescence response style (DRS), a tendency to disagree to statements, and midpoint response style (MRS), a preference for the middle category of a response scale. Response styles are pervasive in questionnaires with Likert-type response scales. For example, Rost, Carstensen, and von Davier (1999) showed the occurrence of ERS and NERS in the German NEO-FFI (Borkenau & Ostendorf, 1993), a finding that was confirmed by Austin, Deary, and Egan (2006) for the English NEO-FFI (Costa & McCrae, 1992). Furthermore, Buckley (2009) investigated the occurrence of ERS, ARS, DRS, and noncontingent responding (random or careless responding; Baumgartner

& Steenkamp, 2001) in attitude scales from the PISA 2006 student questionnaire (OECD, 2006). Other studies examined response styles in a leadership performance scale (Eid & Rauber, 2000), marketing scales on consumer behavior (Baumgartner & Steenkamp, 2001), a measure of tobacco dependence (Bolt & Johnson, 2009), and a survey on cooking behavior (van Herk, Poortinga, & Verhallen, 2004), to name a few.

A number of studies have examined the relationship between response styles and demographic variables and personality traits. In the NEO-FFI, ERS was positively associated with extraversion and conscientiousness (Austin et al., 2006). In the same study women and younger respondents were more likely to employ ERS compared to men and older respondents. The relationship between gender and ERS was also found by Berg and Collier (1953) and Eid and Rauber (2000), though Greenleaf (1992b) and Naemi, Beal, and Payne (2009) did not find any gender differences in the use of ERS. Naemi et al. showed that ERS was related to intolerance of ambiguity, simplistic thinking, and decisiveness. Research on the relationship between ERS and cognitive ability has yielded inconclusive results. Light, Zax, and Gardiner (1965) and Das and Dutta (1969) found that participants of lower intelligence endorsed more extreme responses compared to participants of higher intelligence though Naemi et al. did not find any differences between cognitive ability groups with respect to the use of ERS. Several studies indicate that the use of response styles may differ cross-culturally (Buckley, 2009; Hui & Triandis, 1989; Johnson, Kulesa, Cho, & Shavitt, 2005). For example, Johnson et al. found that ERS was positively related to Hofstede's (2001) cultural dimensions of power distance and masculinity while ARS was negatively related to the same dimensions. Thus, participants from cultures high on power distance and/or masculinity (e.g., Mexico, high on both, or Germany, high on masculinity) may be more likely to employ ERS and less likely to employ ARS compared to participants from cultures low on power distance and/or masculinity (e.g., Australia, low on power distance, or Belgium, low on masculinity).

Response styles can be understood as categorical variables (i.e., a response style is either present or not; e.g., Austin et al., 2006) or as continuous variables that are distributed along a dimension (i.e., people differ in the extent to which they show a certain response style; e.g., Greenleaf, 1992a). The perspective taken has implications for the modeling of response styles: Analyses conducted in the framework of the categorical approach often apply mixed Rasch models to identify latent classes that are assumed to differ qualitatively in their response scale use (Austin et al., 2006; Eid & Rauber, 2000; Rost et al., 1999). Analy-ses following the dimensional approach incorporate a response style dimension in addition to the trait dimension(s) in a multidimensional item response model (Bolt & Johnson, 2009; Bolt & Newton, 2011). Both approaches are implemented in this dissertation and will be contrasted in the discussion.

Trait scores based on summing item responses can be distorted by response styles (Austin et al., 2006; Baumgartner & Steenkamp, 2001; Bolt & Johnson, 2009). For instance, a person employing ERS might receive a more extreme trait score than a person employing NERS even though both have the same latent trait level. Trait estimates derived from mixed Rasch models with latent classes interpretable as response styles or trait estimates from multidimensional item response models that model traits and response styles simultaneously can provide a solution to the distortion of trait scores by correcting trait estimates for response style effects (Bolt & Newton, 2011; Rost et al., 1999). Thus, when response styles play a role, inferences on trait levels and trait comparisons should only be conducted using trait estimates from item response theory (IRT) models that take response styles into account. This important issue will be addressed in detail in the discussion.

The second factor affecting measurement invariance investigated here is the measurement invariance of items between subgroups of respondents. On the side of the items, trait

inferences can be distorted when items are not invariant between different subgroups of respondents, for instance, gender, age, or cultural subgroups. A lack of measurement invariance in this sense is called *differential item functioning* (DIF). An item shows DIF when members of distinct groups differ in the probability of endorsing an item despite having the same latent trait level (Holland & Wainer, 1993). DIF may bias trait scores since the scores of the group it favors may be increased relative to the scores of the disadvantaged group (Reise, Smith, & Furr, 2001). DIF can often be attributed to multidimensionality, i.e., the differentially functioning items measure a secondary dimension in addition to the primary dimension of interest (Shealy & Stout, 1993). When this is the case, item responses do not only depend on the primary dimension but also on the secondary dimension captured by the items. It follows that group differences on the primary dimension then cannot be interpreted as valid differences between the two groups since participants from the two groups may also differ on the secondary dimension. DIF may introduce bias in trait scores and thus systematically favor or disfavor members from a certain subgroup which is problematic when important decisions such as employment decisions or decisions concerning college admission depend on comparisons between individuals based on trait scores. This is why DIF analyses form a critical part of the test validation process in terms of ensuring the instrument's fairness to all respondents according to the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999).

In the ability domain, testing for DIF has been established as a standard practice in test development and validation. However, few applications of DIF research to the personality domain exist. For instance, Mitchelson, Wicher, LeBreton, and Craig (2009) showed that 17 of the 45 scales in the Abridged Big Five Circumplex contained gender-DIF. Smith and Reise

(1998) tested the Multidimensional Personality Questionnaire Stress Reaction Scale for gender-DIF. Reise et al. (2001) found that several items of the NEO-PI-R's (Costa & McCrae, 1992) neuroticism scale, especially on the anxiety facet, showed gender-DIF. This dissertation extends Reise et al.'s research by testing all NEO-PI-R scales for gender-DIF.

Measurement invariance can also refer to the invariance of items across assessments (e.g., of the same instrument at two points in time) which is the third factor pertaining to measurement invariance investigated in this dissertation. Invariance across assessments is especially relevant for trend analyses in large-scale assessments such as PISA (OECD, 2010). Trend analyses investigate the development of student achievements across assessment periods and are for example used to monitor educational reforms. For example, trend analyses may aim at discovering whether the proportion of low-achieving students has increased or decreased. *Link items* (i.e., items that are retained across assessments) have the purpose of ensuring the comparability of scores from different assessments. For example, after reading was the major domain in PISA 2000, 28 of the originally 129 reading items were re-administered in PISA 2003, PISA 2006, and PISA 2009 for linking purposes. The measurement invariance of these link items is an important pre-requisite for trend analyses (Mazzeo & von Davier, 2009) since valid inferences concerning changes in student achievement can only be drawn when the scores are comparable across assessments.

1.3. Papers in this dissertation and their role in investigating the three factors affecting measurement invariance

This dissertation is based on five papers with the following titles and abbreviated designations:

1. Wetzel et al. (in press; DIF)

   Wetzel, E., Böhnke, J. R., Carstensen, C. H., Ziegler, M., & Ostendorf, F. (in press). Do individual response styles matter? Assessing differential item functioning for men and women in the NEO-PI-R. *Journal of Individual Differences*.

2. Wetzel et al. (2013; Consistency response styles)

   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

3. Wetzel & Carstensen (2013b; MIRT)

   Wetzel, E., & Carstensen, C. H. (2013b). *Multidimensional modeling of response styles*. Manuscript submitted for publication.

4. Wetzel & Carstensen (2013c; Reversed thresholds)

   Wetzel, E., & Carstensen, C. H. (2013c). *Reversed thresholds in the Partial Credit Model – A reason for collapsing categories?* Manuscript submitted for publication.

5. Wetzel & Carstensen (2013a; Linking)

   Wetzel, E., & Carstensen, C. H. (2013a). *Linking PISA 2000 and PISA 2009: Implications of instrument design on measurement invariance.* Manuscript submitted for publication.

The paper investigating response styles and differential item functioning in the NEO-PI-R (Wetzel et al., in press; DIF) addresses the first two factors that can potentially distort

trait inferences and trait comparisons (response styles and measurement invariance between subgroups) in one study. The other papers focus on one of the three factors. The papers Wetzel et al. (2013; Consistency response styles) and Wetzel and Carstensen (2013b; MIRT) address specific aspects of the issue of response styles in depth. The paper Wetzel and Carstensen (2013c; Reversed thresholds) addresses the issue of reversed thresholds in the partial credit model, which is related to the topic of response styles. Lastly, the paper Wetzel and Carstensen (2013a; Linking) addresses the measurement invariance of cognitive items across PISA assessments. In the following, the main results of each of the five papers will be summarized.

1. *Do individual response styles matter? Assessing differential item functioning for men and women in the NEO-PI-R (Wetzel et al., in press; DIF)*

The goal of this paper was to analyze the German NEO-PI-R (Ostendorf & Angleitner, 2004) regarding two factors that can influence item responses in addition to an individual's latent trait level, namely individual differences in responses styles and differential functioning of the items between subgroups of respondents (here concerning gender). In the pre-analyses it was shown using a model fit comparison between a constrained mixed partial credit model (constrained mixed PCM) and a mixed partial credit model that several facets of the NEO-PI-R were not homogeneous regarding the trait being assessed. This means that the latent classes did not only differ in their response scale use but also in the construct that was being measured. Thus, these facets were not included in the subsequent analyses on response styles and DIF.

Respondents systematically differed in their response scale use on the remaining NEO-PI-R facets: some preferred extreme categories (ERS) while others preferred moderate categories (NERS). For two facets (openness to actions and deliberation), a third latent class emerged that was also characterized by an NERS but differed from the first NERS class in its avoidance of the middle (neutral) category. Several items especially on neuroticism, agree-

ableness, and conscientiousness showed gender-DIF, indicating that these items were not measurement invariant for men and women. In total, 24 items were classified as slight to moderate and seven items were classified as moderate to large DIF in the complete sample. Most of the DIF items were found on the facets anxiety, angry hostility (both neuroticism), modesty (agreeableness), achievement striving, and deliberation (both conscientiousness). For the complete sample, the direction of DIF was almost balanced: 17 items favored men and 14 items favored women. DIF results were overall consistent between the complete sample, ERS, and NERS, though some differences in the classification and magnitude of DIF existed. Both response styles and DIF can influence item responses, though in our study the two factors appeared to exert their influence largely independently from each other.

2. *Consistency of extreme response style and non-extreme response style across traits (Wetzel et al., 2013; Consistency response styles)*

The aim of this paper was to investigate the consistency of response styles in two instruments, the German NEO-PI-R and several attitude scales from the PISA 2006 student questionnaire. That is, we examined whether participants employed the same response style across the different traits assessed in an instrument or whether they switched between response styles depending on the trait being assessed. In the pre-analyses, respondents were allocated to the NERS or ERS class according to their highest class membership probability in the constrained mixed PCM for each scale. These manifest class memberships were then used as variables in the second order latent class analyses. The second order latent class analyses yielded a two-class solution for the PISA sample and a four-class solution for the NEO-PI-R sample. In both samples, the largest class was characterized by participants who consistently used NERS across scales. The second class in the PISA sample contained par-ticipants who used ERS consistently as well as participants who switched between NERS and ERS. The NEO-PI-

R sample also yielded one class of participants who could not be allocated to a response style clearly but instead appeared to switch between response styles. The remaining two classes in the NEO-PI-R sample were identified as another class of consistent non-extreme responders (but with higher class membership probabilities compared to the first NERS class) and a class of consistent extreme responders. In sum, the second order latent class analyses showed that for the majority of the participants in both samples (65 to 80%), the response style was consistent across traits.

3. Multidimensional modeling of response styles (Wetzel & Carstensen, 2013b; MIRT)

This paper's goal was to apply an alternative method of modeling response styles using multidimensional item response theory (MIRT) models. Thus, in contrast to the first two papers which followed the categorical approach, this paper takes the dimensional approach to modeling response styles. MIRT models have the advantage of allowing traits and response styles to be modeled simultaneously. We applied multidimensional partial credit models (e.g., Kelderman, 1996) to NEO-PI-R data. Each model included one or more trait dimension(s) consisting of the NEO-PI-R facets as well as one or more response style dimension(s) consisting of ERS, ARS, DRS, or MRS.

Comparisons between unidimensional and multidimensional models showed that response styles (especially ERS) were able to explain unique variance in item responses that was incremental to the variance explained by the latent traits. When two response styles were modeled in addition to the trait, the combination of ERS and MRS led to the largest increment in explained variance. Latent correlations revealed that ERS and MRS appear to be largely trait-independent while ARS and DRS are related to several traits. Using weighted likelihood estimates derived from MIRT models we showed that when response styles are incorporated into the model, trait estimates on the substantive NEO-PI-R facets can be corrected for response

styles. Thus, trait estimates obtained in MIRT models are comparable between persons with different response styles since differences in response styles are partialed out during estimation. However, the corrective effect was shown to depend strongly on the modeling of the response style dimension. When the same items as the ones used for the trait dimension are utilized to indicate the response style, a correction takes place since trait variance can be separated from response style variance in the estimation of the traits. However, when separate item sets are used for the trait and response style dimensions the direct model-based correction fails and a different method such as using post-hoc residualized scores has to be applied.

4. *Reversed thresholds in the Partial Credit Model – A reason for collapsing categories? (Wetzel & Carstensen, 2013c; Reversed thresholds)*

When response styles occur, the probability that one or more response categories are not used is increased. For example, with an ERS, responses in moderate categories are underrepresented. Response categories with low frequencies may lead to reversed thresholds in the PCM (for an extensive theoretical treatise on the topic of reversed thresholds see Adams, Wu, & Wilson, 2012). Researchers often deal with reversed thresholds by collapsing categories (e.g., Austin et al., 2006; Nijsten, Sampogna, Chren, & Abeni, 2006; Rost et al., 1999). The goal of this paper was to address empirically using data from the NEO-PI-R as well as simulated data whether this practice is justified in order to avoid reversed thresholds in the PCM.

Our analyses showed that reversed thresholds do not impair the differentiation between respondents with different trait levels since average weighted likelihood estimates for the five response categories were ordered and increased monotonically despite reversed thresholds. Furthermore, mixed PCMs revealed that reversed thresholds were not a phenomenon bearing on the complete sample but only occurred in subgroups of participants. The simulation study

showed that the ordering of the average weighted likelihood estimates can be used to test whether the response categories are ordered since disordered responses will lead to reversals in the average weighted likelihood estimates. The practice of collapsing categories due to reversed thresholds may be justified with misfitting items or when substantive reasons exist. However, in general, it is problematic to collapse categories since participants who chose different response categories are treated as if they expressed the same trait level and important trait information is consequently lost.

5. *Linking PISA 2000 and PISA 2009: Implications of instrument design on measurement invariance (Wetzel & Carstensen, 2013a; Linking)*

This paper addressed the measurement invariance of items across assessment periods using data from several PISA cycles. We investigated the measurement invariance of the common items and link items in the reading domain between PISA 2000 and PISA 2009 as well as the measurement invariance of a subset of the science link items between PISA 2006 and PISA 2009. Furthermore, the size of the link error was investigated for the different item sets.

Model comparisons resulted in a better fit for the models including an interaction term between item and instrument compared to models containing only a group parameter for the instrument. Hence, overall, measurement invariance was violated for both the reading and the science link items. For reading, this pertained both to the instrument from PISA 2000 administered in 2000 and 2009 as well as the instruments from PISA 2000 and PISA 2009 both administered in 2009. Item level analyses revealed that some of the items showed large differences in item difficulty between assessments. Factors that may have exacerbated differences in item difficulty are changes in item wording and position effects, though a model including a three-way interaction between item, cluster position, and instrument for a subset of the common reading items did not yield a better fit than the simpler model including only a group

parameter for the instrument. The size of the link error was shown to depend on sample size and the number of items included in its computation, with larger samples and more items yielding lower link errors.

1.4. Discussion of main results and outlook for future research

In the following, I will first discuss the methods applied in this dissertation to measure response styles. Second, the categorical and dimensional approaches to modeling response styles will be contrasted. Third, the topic of the distortion of trait scores due to response style effects and the correction of trait estimates in mixed Rasch models and multidimensional item response models will be addressed. Fourth, I will discuss how analyses on response styles and differential item functioning can be incorporated in the latent DIF approach suggested by Samuelson (2008). Fifth, other factors that may influence response styles and sixth, methods to reduce the impact of response styles will be considered. Lastly, aspects pertaining to reversed thresholds in the PCM and to the measurement invariance of items across assessments periods will be discussed.

1.4.1. Measuring response styles

According to Baumgartner and Steenkamp (2001, p. 144), ‘‘the major problem in measuring response styles is not to confound stylistic variance with substantive variance’’. Different methods have been suggested to measure response styles and deal with this problem. These include classical methods such as the computation of indices as well as methods that apply item response models to model response styles as a latent variable. The most straightforward way of computing a response style index is to count the frequency with which certain response categories were endorsed by a respondent (e.g., the extreme cate-gories to measure ERS or categories stating agreement to measure ARS). Solving the problem of not confounding stylistic variance with substantive variance is then attempted by using an item set composed of heterogeneous items with low inter-item correlations to compute the response style index (Baumgartner & Steenkamp 2001; Greenleaf, 1992a). De Beuckelaer, Weijters, and Rutten (2010) go further and recommend using a random sample of items from multiple scales

that are not relevant to the construct of interest to assess a person's response style. In the case of measuring ARS and DRS, Baumgartner and Steenkamp suggest balancing the scale with respect to positively and negatively worded items. Due to testing time restrictions and considerations regarding test-taker fatigue, administering a number of items only for the purpose of assessing response styles as advocated by De Beuckelaer et al. is often not feasible. Approaches that model respondents' response style using the same items as the ones assessing the trait of interest are thus preferable in this respect. One such approach is presented by Meiser and Böckenholt (2011; see also Böckenholt, 2012) who distinguish different response processes that take place during the completion of an item: those that are related to the trait and those that are related to response styles. The modeling of the response processes relies on pseudo-items that indicate the decision a respondent made during each sub-process: A decision on 1) whether to endorse the middle category or not (i.e., MRS), 2) the direction of the attitude (agree or disagree), and 3) the intensity of the attitude (non-extreme or extreme response option; i.e., ERS). These response processes can be modeled using a multidimensional item response model. Meiser and Böckenholt found that this multidimensional model described the data better than a unidimensional model which they argue means that the multidimensional model succeeded at differentiating trait processes from response style processes.

This dissertation applied two item response methods to measure response styles which also treat response styles as latent variables but are more direct and do not require the construction of pseudo-items, namely mixed Rasch models and multidimensional item response models. These item response models provide two solutions to the problem of distinguishing stylistic variance from substantive variance in the measurement of response styles. First, in mixed Rasch models, persons that differ systematically in their response patterns with regard to a preference or avoidance of extreme categories can be allocated to separate latent classes that consist of distinct response styles. Within each of these latent classes quantitative trait

differences can be examined. To ensure that the latent classes only differ regarding response styles and stylistic variance is thus separated from substantive variance, a constraint should be implemented in the mixed Rasch model that restricts item location parameters to be equal between latent classes. If this constrained model describes the data better than a model in which all parameters are estimated freely, it can be deduced that the latent classes capture only stylistic variance and are not confounded with substantive variance (see Wetzel et al., 2013; Consistency response styles).

Second, the problem of separating stylistic variance from substantive variance in measuring response styles (Baumgartner & Steenkamp, 2001) can be solved by estimating MIRT models that incorporate several traits and one or more response styles. With only one trait and one response style it is hard to differentiate whether a person in fact has a very high latent trait level or whether he or she is, for example, an extreme responder. If a high amount of extreme responses is consistently given across several traits, it is unlikely that the person has a very high latent trait level on all these traits. Instead, then it can be concluded that he or she is an extreme responder. Thus, in MIRT models stylistic variance can be separated from substantive variance by including a dimension that represents the response style in addition to the traits of interest. For this response style dimension item responses are coded differently from the trait dimensions. For example, for ERS only extreme responses are scored with 1 while the other categories are scored with 0. For ARS, response options stating agreement are scored with 1 while response options stating disagreement or neutrality are scored with 0. However, the success of separating the two types of variance also depends on the specific response style being measured. In the former case of ERS the scoring of item responses to indicate ERS and the traits implies that these dimensions are independent. Zero or low correlations between ERS and the trait dimensions confirm that they are largely independent. For the latter, namely ARS,

stylistic variance and substantive variance may still be partly confounded in the ARS dimension as indicated by the dependency in the scoring of item re-sponses and by medium-sized correlations between ARS and many NEO-PI-R facets (Wetzel & Carstensen, 2013b; MIRT). The operationalization of the response style dimension in MIRT models (i.e., based on the same items as the trait dimensions or based on items from different traits) is addressed in Wetzel and Carstensen (2013b; MIRT) and poses an issue that warrants further investigation.

1.4.2. Categorical and dimensional approaches to modeling response styles

Both the categorical approach using mixed Rasch models and the dimensional approach using MIRT to modeling response styles were applied in this dissertation. Can either of the approaches be evaluated as the more appropriate approach? On the one hand, mixed Rasch analyses indicated that qualitative differences exist between the latent classes identified as distinct response styles. On the other hand, our investigation of the consistency of NERS and ERS across the traits assessed in an instrument indicates that quantitative differences in response styles may prevail. We confirmed previous findings on the consistency of response styles (e.g., Austin et al., 2006; Hernández, Drasgow, & González-Romá, 2004) using a new method, namely second order latent class analysis. In our study 65 to 80% of the respondents applied the same response style independently of the trait being measured. However, this also means that between 20 and 35% of the respondents could not be allocated to either NERS or ERS clearly, indicating that they switched between NERS and ERS several times over the course of the instrument. Fluctuations in the percentage of respondents being classified as consistent response style users vs. “switchers” may be due to differences across samples (heterogeneous population sample vs. 15-year old students) and instruments (NEO-PI-R vs. PISA attitude scales) as well as uncertainty in the estimation of class membership. The use of NERS and ERS did not appear to be systematic or contingent on the trait being assessed for

the class of "switchers" since membership probabilities to NERS and ERS were between about 40 and 60% for all traits in the second order latent class analysis. It seems important to investigate the characteristics of this class of "switchers" further and to elucidate trait and situational factors associated with the consistency or inconsistency of response styles. For example, studies could attempt to identify covariates linked to membership in one of the latent classes and try to predict class membership with these covariates. Possible covariates could be demographic variables or trait variables associated with NERS or ERS (e.g., gender or intolerance of ambiguity; see Austin et al., 2006; Naemi et al., 2009). Situational factors that may play a role in influencing the consistency of response styles could pertain to the testing situation, such as whether it is a high-stakes or low-stakes situation and the respondents' level of anonymity.

Overall, differences in membership probabilities to NERS or ERS in the latent classes obtained in the second order latent class analyses appeared to be mainly quantitative in nature, implying that NERS and ERS might be poles of the same dimension. The consistency of response styles across traits for the majority of respondents raises the question of whether response styles can be understood as a latent trait or whether they are "nuisance" (Bolt & Newton, 2011). Relationships of response styles to substantive traits (e.g., Austin et al., 2006; Wetzel & Carstensen, 2013b; MIRT) and the longitudinal stability of ERS, ARS, DRS, and MRS across a 1-year period (Weijters, Geuens, & Schillewaert, 2010) suggest that response styles may also be of substantive interest and may in part be "a reflection of "real" and stable traits" (Cronbach, 1950, p. 16-17). Furthermore, response styles differ from other response biases such as faking or dissimulation in that they are not intentional behavior that occurs only on scales judged to be relevant by the respondent (e.g., assessing attributes needed for a job) but instead occur on all scales (MacCann, Ziegler, & Roberts, 2012). If future research supports the notion of response styles as latent trait variables, this would be an argument in favor of the dimensional approach.

One disadvantage of the categorical approach is that when latent classes are not homogeneous concerning the trait (indicated by the mixed PCM showing a better fit than the constrained mixed PCM), differences between the latent classes cannot be attributed to response styles with certainty. Scales where this is the case should be removed when response styles are the study object. In contrast, the dimensional approach allows analyzing all scales since here trait heterogeneity and heterogeneity in response scale use can be modeled simultaneously. Furthermore, there is a substantial amount of variation in the membership probabilities estimated in mixed constrained PCMs indicating that the allocation of respondents to one response style group involves a certain degree of uncertainty (though overall membership probabilities were high (>.80) in our analyses). Moreover, in mixed Rasch models participants who do not have a distinct response style but instead use the response scale evenly are generally assigned membership in the NERS class since their membership probability for this class is highest. However, it is questionable whether this is an adequate manner of treating this group of respondents. In multidimensional models this problem does not exist since each respondent's individual degree of preferring or avoiding extreme responses is modeled accurately.

The dimensional approach allows investigating the relative influence of different response styles whereas the categorical approach using mixed Rasch models only appears to be able to differentiate between NERS and ERS while other response styles such as ARS and DRS were not found in the latent classes derived from mixed Rasch models. In our analyses, multidimensional models containing a response style dimension showed a superior fit compared to unidimensional models and ERS was the response style that explained the largest amount of variance incremental to the trait. Since ERS appears to be the most important response style it could be argued that – in the interest of parsimony – differentiating between NERS and ERS might be sufficient. However, the prevalence of ERS would have to be vali-

dated using different instruments and samples, especially samples from different cultural backgrounds than the German samples used here. Considering that for example Buckley (2009) and Johnson et al. (2005) showed that cross-cultural differences in response styles exist, it seems important to conduct cross-cultural analyses to investigate whether in other cultures (that are for example lower on individualism than Germany) different response styles (e.g., acquiescence) play a more important role. In sum, the research reported in this dissertation indicates that the dimensional approach yields important advantages over the categorical approach. Many of these advantages can be ascribed to its flexibility in modeling different types of response styles and in modeling inter-individual differences in response styles. Nevertheless, further research could compare the two approaches systematically and investigate under which conditions one might be more appropriate than the other.

1.4.3. Correction of trait estimates for response style effects

Related to the comparison of the categorical and dimensional approaches is the issue of obtaining trait estimates that are corrected for response style effects. When response styles occur, trait scores that are based on summing the responses to all items in a scale can be biased. Depending on whether the scale is balanced, ERS-respondents may receive more extreme trait scores than NERS-respondents (Austin et al., 2006). Furthermore, the magnitude of the bias depends on the person's trait level: at average trait levels the bias is smaller than at very low or very high trait levels (Bolt & Newton, 2011). Hence, persons with different response styles are not directly comparable when trait scores are used. The same is the case with trait estimates (e.g., weighted likelihood estimates) derived from unidimensional PCMs that do not take response styles into account. Wetzel et al. (2013; Consistency response styles) showed for ERS that correlations between personality traits were higher in the standard unidimensional PCM compared to the constrained mixed PCM. Since response style effects are partialed out in the

constrained mixed PCM, this indicates that systematic variance due to ERS may have increased correlations in the unidimensional PCM (see also Austin et al., 2006). Using residualized scores, Baumgartner and Steenkamp (2001) also found that trait correlations were overestimated due to response style effects. These results imply that response styles may impact the construct validity of instruments. Correlations taken as evidence for the convergent and discriminant validity of an instrument may be over-estimated due to common response style variance. Future studies could clarify the impact of response styles on different forms of validity.

The two item response models used within the categorical and dimensional approaches, namely mixed Rasch models and MIRT models, can both lead to corrected trait estimates that account for the respondents' response style. In the case of mixed Rasch models, members of the ERS class receive less extreme trait estimates whereas trait estimates for members of the NERS class are adjusted to be more extreme (Wetzel et al., 2013; Consistency response styles). Thus, participants with the same sum score will receive different trait estimates, depending on whether they are allocated to the ERS or NERS class.

In MIRT models, both traits and response styles are modeled simultaneously, allowing a person's level on, for example, ERS to be taken into account in the estimation of his or her trait level. Wetzel and Carstensen (2013b; MIRT) showed that weighted likelihood estimates from unidimensional PCMs did not correlate perfectly with those from two-dimensional PCMs that included an ERS dimension, indicating that the trait estimates in two-dimensional models were adjusted for the respondents' level of ERS. Furthermore, while sum scores on ERS were distributed evenly over sum scores on personality traits, weighted likelihood estimates on ERS were negatively related to weighted likelihood estimates on the respective trait in two-dimensional models, showing that an adaptation for the respondents' ERS took place. However, whether a correction for response style effects is implemented in MIRT models depends

strongly on how the response style dimension is modeled, that is, whether the same items are used to model both the response style and trait dimensions (correction occurs) or whether two separate item sets are used (no correction). When the same items are used response style effects are accounted for in the estimation of trait variance and, in consequence, trait estimates are corrected for response style effects. However, when separate item sets are used, trait variance and response style variance are confounded in the estimation of trait variance, resulting in trait estimates that are contaminated by response style effects.

The mechanisms underlying the corrective effect observed in mixed Rasch models and MIRT models need to be investigated further. For example, it would be important to understand which method yields trait estimates that are closest to the true latent trait levels. It could be hypothesized that the correction might work better in the application of MIRT models since the individual level of, for instance, ERS is taken into account, allowing for a more precise adaptation of trait estimates. In contrast, mixed Rasch models only differentiate between ERS and NERS and do not allow for different levels of ERS or NERS. The correction applied to trait estimates should then be the same for all respondents in one class. Alternative approaches should also be considered in future studies investigating methods of correcting for response styles. For example, Baumgartner and Steenkamp (2001) advocate the use of residualized scores, a method that can be applied with classic response style indices. For instance, a simulation study could be conducted that compares the different methods of correcting for response styles with respect to their ability of recovering the true latent trait levels.

1.4.4. Response styles, differential item functioning, and the latent DIF approach

In our study on response styles and gender-DIF we conducted DIF analyses in the complete sample as well as separately within response style groups. Since DIF may be due to

multidimensionality (Shealy & Stout, 1993), the goal of this procedure was to control for differences in response styles as the potential secondary dimension which may have been responsible for DIF. However, both response styles and gender-DIF appeared to influence item responses independently in our data since DIF results were overall consistent between the complete sample and the NERS and ERS subsamples. This may be attributed to the small relationship between membership to NERS or ERS and gender for some NEO-PI-R facets and the non-significant relationship for other facets. Future studies could investigate the influence of response styles on DIF analyses when response styles are related to the group variable of interest used in DIF analyses. Bolt and Johnson's (2009) analysis of DIF between a low-education group and a high-education group in a measure of tobacco dependence suggests that when a relationship between response styles and the group variable exists (in this case higher ERS in the low-education group), DIF can partly be attributed to response style effects. Possible group variables besides education might be related to culture (e.g., Johnson et al., 2005), ethnicity (Bachman & O'Malley, 1984), intolerance of ambiguity, or simplistic thinking (Naemi et al., 2009).

The latent approach to DIF proposed by Samuelson (2008), which operates without a manifest group variable but instead differentiates groups based on a latent class analysis, is also a promising area of future research that can be linked to the research in this dissertation. In Samuelson's approach, differential item functioning is investigated between the latent classes derived from a latent class analysis. The latent DIF approach overcomes some of the problems associated with the classic (manifest) DIF approach, namely the lack of homogeneity within the manifest groups and the lack of a relationship between the manifest groups and the trait of interest (Samuelson, 2008). By applying mixed Rasch models, the latent DIF approach ensures that latent classes differ maximally from each other while homogeneity exists within latent classes (Rost, 1990). For cognitive tests the latent classes can be used directly in DIF

analyses since response styles are not an issue, though individual differences in dealing with the speed aspect of the test (e.g., guessing, random responding toward the end of the test) might influence the number and nature of the emerging latent classes. For constructs commonly assessed with Likert-type scales (e.g., personality, interests, and attitudes), the latent DIF approach cannot be applied in the same straightforward manner since the latent classes may differ regarding response styles, the construct being measured, or other factors. Nevertheless, the latent DIF approach suggested by Samuelson (2008) could be extended to the context of multi-categorical items in questionnaires by applying the model comparison between a constrained mixed PCM and a mixed PCM introduced in Wetzel et al. (2013; Consistency response styles) as a first step. This would ensure a separation between response style variance and variance in item responses that can be attributed to a group variable or differences in the construct being assessed. If the mixed PCM shows a better fit, heterogeneity in the latent classes exists with respect to multiple factors, among them response styles and the trait. Since different sources of variance are confounded in this case, the latent classes derived from the mixed PCM should not be used for latent DIF analyses. The latent classes emerging in the constrained mixed PCM can only differ regarding response styles due to the constraint implemented in the model. Thus, if one is interested in DIF between response style groups, DIF can be analyzed between these latent classes. In contrast, if one is interested in analyzing DIF between latent classes that differ in other factors besides response styles, it would be necessary to conduct a latent class analysis within the response style groups obtained in the constrained mixed PCM and to then analyze DIF between these new latent classes. While this approach admittedly is somewhat complicated and involves several steps, it achieves a clean separation between response styles and heterogeneity based on the trait and other factors which is essential where response styles are an issue.

Analyses on differential item functioning cover one aspect of measurement invariance and play an important role in test validation (American Educational Research Association et al., 1999). Differentially functioning items can imply item bias in terms of one subgroup of respondents being favored over another. For example, Wetzel et al. (in press; DIF) analyzed data from the German NEO-PI-R and found that some items especially on neuroticism, agreeableness, and conscientiousness facets showed DIF for men and women and therefore may contain a bias. Since questionnaires such as the NEO-PI-R are often applied in personnel selection or other high-stakes situations with far reaching consequences for individuals, it is critical to ensure the fairness of the instrument to all respondents. Samuelson's (2008) latent DIF approach offers a framework for DIF studies that overcomes some of the disadvantages of using manifest groups. Incorporating it with other factors that may influence measurement invariance such as response styles would be an important endeavor for future research.

1.4.5. Further factors influencing response styles

Further factors that may influence the occurrence and intensity of response styles should be considered. Among these are the number of negatively worded items on a scale and the social desirability of the items. Baumgartner and Steenkamp (2001) state that using negatively worded items can reduce the influence of ARS and DRS though it does not counteract ERS. However, according to Marsh (1996), the disadvantages brought about by using negatively worded items (i.e., method effects that lead to the emergence of a second factor) may outweigh the advantages of counteracting response styles. In our analyses, correlations between the NEO-PI-R facets and ARS were stronger for increasing numbers of negatively worded items, though this relationship was not found for other response styles. With more negatively worded items it is more likely that some participants (especially participants with lower verbal ability; Marsh, 1986, 1996) will employ ARS. Considering that ARS is defined

as stating agreement irrespective of item content, positive relationships to traits could be expected to decrease when ARS is employed more often with increasing numbers of negatively worded items, though negative relationships between ARS and traits might grow stronger. The confounding of trait-response style relationships with the number of negatively worded items is an issue that deserves further attention. It could for example be investigated by systematically manipulating the number of negatively worded items on scales assessing the same construct.

Socially desirable responding, i.e., "the tendency to give overly positive self-descriptions" (Paulhus, 2002, p. 50) often occurs on personality items (e.g., Bäckström, Björklund, & Larsson, 2009). This raises the question of how different response biases such as socially desirable responding and response styles are related. Further research could investigate the effects of the social desirability of the items on the occurrence of response styles and on the relationships between response styles and traits. For example, it would be interesting to find out whether the relationships between traits and ARS or DRS are diminished when social desirability is partialed out and to compare the proportion of respondents that are classified as extreme responders or acquiescent responders between a model that takes social desirability into account and one that does not. Here, MIRT models are also a feasible method since an additional dimension that represents the social desirability of the participants' responses could be added to the trait dimension and the response style dimension.

Another factor that might exert an influence on the occurrence and intensity of response styles is the topic diversity of the items (Weijters, Geuens et al., 2010). It is conceivable that redundancy in item content may exacerbate response styles since respondents' inclination towards careless responding might increase. In connection with this it would also be interesting to investigate how response styles are related to motivational variables such as test-taking motivation and contextual variables associated with the testing situation. For example, in high-

stakes situations such as personnel selection it could be assumed that less careless responding (without regard to item content) takes place. This could be investigated by comparing a sample with a regular instruction with a sample with an instruction that constructs an applicant setting, though in this situation methods would have to be developed that can distinguish response styles from faking and socially desirable responding.

The use of response styles may also be related to personality traits. For the Big Five several studies have investigated this. For example, Austin et al. (2006) found a positive relationship between ERS and extraversion and conscientiousness in the NEO-FFI, though ERS was largely independent of the NEO-PI-R facets (Wetzel & Carstensen, 2013b; MIRT). Relationships between response styles and other personality traits that go beyond the Big Five would be an interesting topic of research. One study by Naemi et al. (2009) found that ERS was associated with intolerance of ambiguity, simplistic thinking, and decisiveness. ERS also appears to be positively related to anxiety (Berg & Collier, 1953; Lewis & Taylor, 1955). Other personality constructs that might be of interest with respect to response styles could be need for cognition and boredom susceptibility since these might influence how a participant deals with the rather monotone task of completing a questionnaire. Participants high on need for cognition could be expected to expend the necessary cognitive effort and complete the questionnaire meticulously (negative relationship with response styles) whereas participants high on boredom susceptibility might be more tempted to disregard item content and rush through the questionnaire (positive relationship with response styles).

1.4.6. Reducing the impact of response styles

Using trait estimates derived from mixed Rasch models or multidimensional models when response styles pose an issue is a method of dealing with response style effects post-hoc.

However, measures can also be taken to reduce the occurrence and impact of response styles beforehand. Future research could explore ways of reducing the occurrence of response styles that are aimed at the questionnaire itself such as the response format, the labeling of the response options, the cognitive load it imposes, or the instruction. For example, Weijters, Cabooter, and Schillewaert (2010) deduced from their analyses that using a fully labeled scale and a scale that includes a neutral point reduced ERS though ARS increased when all response categories received a label. Concerning the number of response categories it might be advisable to use an even number of categories since MRS and other problems associated with the middle category (e.g., Hernández et al., 2004) could be avoided this way. Results concerning the number of response categories are mixed: According to Kieruj and Moors (2011), ERS is not affected by scale length though Cronbach (1946) recommended reducing the number of response options to counter response styles. The latter contention is also supported by Naemi et al.'s (2009) finding that ERS appears to be related to simplistic thinking and intolerance of ambiguity, implying that response scales should not be too differentiated. Thus, participants with higher levels on these two traits might be more inclined to employ an ERS when the response scale is highly differentiated (e.g., eight response options compared to four). The mode of data collection appears to play a role as well: Weijters, Schillewaert, and Geuens (2008) found slightly lower levels of ERS and DRS in online data compared to data collected using telephone interviews and a paper-pencil administration. Using a forced-choice format is another possibility to counter response styles. Brown and Maydeu-Olivares (2011) introduced a multidimensional item response model to analyze forced-choice items in personality assessment which can yield normative data. Their approach is a promising avenue of further research since it can eliminate some response styles such as ARS.

Test developers could also attempt to formulate items that elicit response styles to a lesser degree. For instance, ambiguous or unstructured items appear to intensify ARS since

response styles tend to occur when respondents are uncertain about item content (Cronbach, 1946). Bäckström et al. (2009) showed that rephrasing personality items to make them more neutral reduced socially desirable responding. This approach might also be feasible for reducing the occurrence of response styles. A high cognitive load appears to increase ARS (Knowles & Condon, 1999), implying that the items and the testing situation should be evaluated regarding the cognitive load they impose. The latter is in line with research on *optimizing* vs. *satisficing* in the response process (Krosnick, 1999). Optimizing means that participants give optimal answers in the sense of executing all the cognitive processes involved in the response process and expending the necessary cognitive effort. In this case response styles should not occur. However, when respondents satisfice (i.e., they either execute the response process less diligently or even skip steps in the response process), response styles such as ARS or MRS result. To reduce satisficing, Krosnick recommends taking measures to for example lower the task difficulty and increase the motivation of participants.

In addition, the questionnaire's instruction could attempt to sensitize respondents for response styles and stress the importance of sincere responses to the validity of the researcher's results. This might lead to a more even use of the response scale though whether this works would have to be verified in a study that compares the occurrence of response styles between different instructions. Other response biases, especially socially desirable responding, have also been investigated using behavioral data and physiological data (see Uziel, 2010 for a review). Thus, research aimed at reducing the impact of response styles could investigative behavioral or physiological indicators of response styles.

Another way of circumventing response style effects or of reducing the impact of response styles by validating self-report data would be to use data based on other-ratings. Other-ratings of Big Five traits for example have incremental validity over self-ratings of Big Five traits and cognitive ability in predicting academic performance in high school students

(Bratko, Chamorro-Permuzic, & Saks, 2006) and university students (Ziegler, Danay, Schölmerich, & Bühner, 2010). Thus, other-ratings might provide further information on the true latent trait levels of respondents that are unbiased by the person of interest's response style though, presumably, the respondent giving the other-ratings may also employ a response style.

1.4.7. Reversed thresholds in the partial credit model

The topic of reversed thresholds and collapsing categories to deal with them is relevant to response styles, since response styles facilitate categories with low frequencies which lead to reversed thresholds, but also in general to questionnaires applying Likert-type scales. From a theoretical perspective, response categories should not be collapsed due to reversed thresholds in the partial credit model because reversed thresholds do not violate model assumptions (Adams et al., 2012; though Andrich, 2013, argues differently). From an empirical perspective, categories should not be collapsed solely due to reversed thresholds for three reasons: 1) Trait averages per category are ordered along the latent trait despite reversed thresholds, 2) reversed thresholds often only occur in subgroups of respondents, and 3) response categories differentiate between respondents with different trait levels, indicating that trait information is lost when categories are collapsed.

Thus, collapsing categories should not be an automatism when reversed thresholds occur but should be justified with other reasons besides reversed thresholds. For example, categories with low frequencies may indicate that the response scale was too differentiated for respondents or that the items were too easy or too difficult. Other reasons may pertain to item properties such as misfit or multidimensionality in the items. Reversed thresholds may also imply that the assumption of equal item discriminations in the PCM may not be adequate (Adams et al., 2012) and that a model which allows differing item discriminations (e.g., the

generalized PCM; Muraki, 1992) might provide a more appropriate description of the response data.

1.4.8. Measurement invariance of items across assessment periods

The measurement invariance of items across assessment periods is especially important with longitudinal test designs or large-scale assessments that conduct trend analyses since in these applications the validity of comparisons across assessments relies on the relationship between the scores and the construct they represent staying the same. For example, in longitudinal test designs that monitor treatment outcomes in clinical practice, patients are repeatedly administered a self-report instrument over the course of the treatment. Here, the validity of changes in the patients' scores depends on the measurement invariance of the items across assessment intervals. With repeated measurements of the same individuals, response shifts (i.e., changes in the respondents' standards for measurement) may threaten measurement invariance and thus confound comparisons (Fokkema, Smits, Kelderman, & Cuijpers, 2013).
This dissertation investigated the measurement invariance of items across assessment periods in a large-scale assessment, namely PISA, where trend analyses are conducted using scores from different cohorts of students across assessments. To achieve measurement invariance across assessments, our results suggest that rather more items should be used as link items and that items with large differences in item difficulty between assessments should be removed from the link. Both recommendations reduce the link error and thus contribute to stable trends. Furthermore, the position of items within clusters should be maintained since position effects may contribute to differences in item difficulty between assessments. Wu (2010) also notes that administration conditions should remain the same in order to reduce the link error. These recommendations are in line with Mazzeo and von Davier's (2009, p. 4) stance of assuming that "all changes potentially matter and should be avoided where possible" ". Country-DIF

(i.e., an interaction between items and countries) further threaten trend analyses. Thus, the stability of trends can also be improved by applying test designs and sampling techniques that reduce the link error.

1.5. Concluding remarks

Establishing measurement invariance is essential for drawing accurate and valid inferences concerning respondents' trait levels and for conducting comparisons between respondents from different groups. This dissertation investigated three aspects of measurement invariance in detail, namely 1) individual differences in response styles, 2) the measurement invariance of items between subgroups of respondents, and 3) the measurement invariance of items across assessment periods. If measurement invariance is violated regarding these three factors, trait scores do not represent the construct of interest purely. Instead, the measurement of the construct is contaminated by one or several secondary dimensions that influence item responses in addition to the respondent's latent trait level. The findings of this dissertation emphasize that individual differences in response styles play an important role in questionnaire data since they explain variance in item responses that is incremental to the variance explained by the trait. This implies that trait scores may be biased and in consequence trait inferences and comparisons between individuals and groups based on sum scores can lead to distorted conclusions. However, response styles appear to be largely consistent across scales and there is tentative evidence that they might also be stable over time. This raises the possibility of correcting trait scores for response style effects. Model-based approaches such as mixed Rasch models or multidimensional item response models are able to separate stylistic variance from substantive variance in the measurement of response styles and to separate variance due to response styles from trait variance. This allows them to provide trait estimates that are corrected for response style effects. Therefore, trait comparisons should only be conducted using trait estimates containing a model-based correction for response style effects.

Another implication of the pervasiveness of response styles in questionnaire data is that response style effects should be taken into account in the process of test validation since correlations between traits that are taken as evidence for construct validity may be spuriously

increased due to common response style variance. Thus, steps should be taken to minimize the influence of factors that can interfere with measurement invariance. A number of measures can be taken in test construction to reduce their occurrence and impact, for example pertaining to item properties, response format, and test motivation. Nevertheless, many open research questions remain. Concerning response styles these include the influence of the social desirability of the items on response styles, a comparison of different methods of correcting trait estimates for response style effects, and properties of the instrument that can reduce the occurrence and impact of response styles.

1.6. References

Adams, R. J., Wu, M. L., & Wilson, M. (2012). The Rasch rating model and the disordered threshold controversy. *Educational and Psychological Measurement, 72*(4), 547–573. doi:10.1177/0013164411432166

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Education Research Association.

Andrich, D. (2013). An expanded derivation of the threshold structure of the polytomous Rasch Model that dispels any "threshold disorder controversy". *Educational and Psychological Measurement*, *73*(1), 78–124. doi:10.1177/0013164412450877

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

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. doi:10.1086/268845

Bäckström, M., Björklund, F., & Larsson, M. R. (2009). Five-factor inventories have a major general factor related to social desirability which can be reduced by framing items neutrally. *Journal of Research in Personality*, *43*(3), 335–344. doi:10.1016/j.jrp.2008.12.013

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

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

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

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. doi:10.1177/0146621608329891

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

Borkenau, P., & Ostendorf, F. (1993). *NEO-Fünf-Faktoren-Inventar (NEO-FFI) nach Costa und McCrae*. Göttingen, Germany: Hogrefe.

Borsboom, D., Romeijn, J.-W., & Wicherts, J. M. (2008). Measurement invariance versus selection invariance: Is fair selection possible? *Psychological Methods*, *13*(2), 75–98. doi:10.1037/1082-989X.13.2.75

Bratko, D., Chamorro-Premuzic, T., & Saks, Z. (2006). Personality and school performance: Incremental validity of self- and peer-ratings over intelligence. *Personality and Individual Differences*, *41*(1), 131–142. doi:10.1016/j.paid.2005.12.015

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

Buckley, J. (2009). *Cross-national response styles in international educational assessments: Evidence from PISA 2006*. Retrieved January, 2012 from https://edsurveys.rti.org/PISA/.

Costa, P. T., & McCrae, R. R. (1992). *Revised NEO personality inventory (NEO-PI-R) and NEO Five-Factor inventory (NEO-FFI)*. Odessa, FL: Psychological Assessment Resources.

Cronbach, L. J. (1946). Response sets and test validity. *Educational and Psychological Measurement*, *6*, 475–494. doi:10.1177/001316444600600405

Cronbach, L. J. (1950). Further evidence on response sets and test design. *Educational and Psychological Measurement*, *10*(1), 3–31. doi:10.1177/001316445001000101

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

De Beuckelaer, A., Weijters, B., & Rutten, A. (2010). Using ad hoc measures for response styles: A cautionary note. *Quality & Quantity*, *44*(4), 761–775. doi:10.1007/s11135-009-9225-z

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

Fokkema, M., Smits, N., Kelderman, H., & Cuijpers, P. (2013). Response shifts in mental health interventions: An illustration of longitudinal measurement invariance. *Psychological Assessment.* doi:10.1037/a0031669

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*(b), 328–351. doi:10.1086/269326

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

Hofstede, G. H. (2001). *Culture's consequences: Comparing values, behaviors, institutions, and organizations across nations* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum.

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

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. doi:10.1177/0022022104272905

Kelderman, H. (1996). Multidimensional Rasch Models for partial-credit scoring. *Applied Psychological Measurement*, *20*(2), 155–168. doi:10.1177/014662169602000205

Kieruj, N. D., & Moors, G. (2011). Response style behavior: Question format dependent or personal style? *Quality & Quantity*, *47*(1), 193–211. doi:10.1007/s11135-011-9511-4

Knowles, E. S., & Condon, C. A. (1999). Why people say "yes": A dual-process theory of acquiescence. *Journal of Personality and Social Psychology*, *77*(2), 379–386. doi:10.1037/0022-3514.77.2.379

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology*, *50*, 537–567.

Lewis, N. A., & Taylor, J. A. (1955). Anxiety and extreme response preferences. *Educational and Psychological Measurement*, *15*(2), 111–116. doi:10.1177/001316445501500203

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. doi:10.1037/h0022746

MacCann, C., Ziegler, M., & Roberts, R. D. (2012). Faking in personality assessment: Reflections and recommendations. In M. Ziegler, C. MacCann, & R. D. Roberts (Eds.),

New perspectives on faking in personality assessment (pp. 309–329). New York, NY: Oxford University Press.

Marsh, H. W. (1996). Positive and negative global self-esteem: A substantively meaningful distinction or artifactors? *Journal of Personality and Social Psychology*, *70*(4), 810–819. doi:10.1037/0022-3514.70.4.810

Marsh, H. W. (1986). Negative item bias in ratings scales for preadolescent children: A cognitive-developmental phenomenon. *Developmental Psychology*, *22*(1), 37–49.

Mazzeo, J. & Davier, M. von. (2009). *Review of the Programme for International Student Assessment (PISA) test design: Recommendations for fostering stability in assessment results*. Retrieved from http://edsurveys.rti.org/PISA/

Meiser, T., & Böckenholt, U. (2011, September). *IRT-Analyse von Traitausprägung und Antwortstilen in Ratingdaten* [IRT analysis of trait levels and response styles in rating scale data], Bamberg, Germany.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research*, *13*, 127–143. doi:10.1016/0883-0355(89)90002-5

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543.

Mitchelson, J. K., Wicher, E. W., LeBreton, J. M., & Craig, S. B. (2009). Gender and ethnicity differences on the Abridged Big Five Circumplex (AB5C) of personality traits: A differential item functioning analysis. *Educational and Psychological Measurement*, *69*(4), 613–635. doi:10.1177/0013164408323235

Muraki, E. (1992). A generalized Partial Credit Model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*(2), 159–176. doi:10.1177/014662169201600206

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

Nijsten, T. E. C., Sampogna, F., Chren, M.-M., & Abeni, D. D. (2006). Testing and reducing Skindex-29 using Rasch analysis: Skindex-17. *Journal of Investigative Dermatology*, *126*(6), 1244–1250. doi:10.1038/sj.jid.5700212

OECD. (2006). *PISA 2006: Assessing scientific, reading and mathematical literacy*. Paris, France: OECD Publications.

OECD. (2010). *PISA 2009: Learning trends: Changes in student performance since 2000*. Paris, France: OECD Publications.

Ostendorf, F., & Angleitner, A. (2004). *NEO-PI-R: NEO-Persönlichkeitsinventar nach Costa und McCrae*. Göttingen, Germany: Hogrefe.

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: Academic Press.

Paulhus, D. L. (2002). Social Desirable Responding: The Evolution of a Construct. In H. I. Braun, D. N. Jackson, D. E. Wiley, & S. Messick (Eds.), *The role of constructs in psychological and educational measurement* (pp. 49–69). Mahwah, NJ: L. Erlbaum.

Reise, S. P., Smith, L., & Furr, R. M. (2001). Invariance on the NEO-PI-R neuroticism scale. *Multivariate Behavioral Research*, *36*(1), 83–110. doi:10.1207/S15327906MBR3601_04

Rost, J. (1990). Rasch Models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, *14*(3), 271–282. doi:10.1177/014662169001400305

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

Samuelsen, K. M. (2008). Examining differential item functioning from a latent mixture perspective. In G. R. Hancock & K. M. Samuelsen (Eds.), *Advances in latent variable mixture models* (pp. 177–197). Charlotte, NC: Information Age Pub.

Shealy, R. T., & Stout, W. F. (1993). A model based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika*, *58*(2), 197–239. doi:10.1007/BF02294572

Smith, L. L., & Reise, S. P. (1998). Gender differences on negative affectivity: An IRT study of differential item functioning on the Multidimensional Personality Questionnaire Stress Reaction scale. *Journal of Personality and Social Psychology*, *75*(5), 1350–1362. doi:10.1037/0022-3514.75.5.1350

Uziel, L. (2010). Rethinking social desirability scales: From impression management to interpersonally oriented self-control. *Perspectives on Psychological Science*, *5*(3), 243–262. doi:10.1177/1745691610369465

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

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. (2010). 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

Wetzel, E., Böhnke, J. R., Carstensen, C. H., Ziegler, M., & Ostendorf, F. (in press). Do individual response styles matter? Assessing differential item functioning for men and women in the NEO-PI-R. *Journal of Individual Differences*,

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., & Carstensen, C. H. (2013a). *Linking PISA 2000 and PISA 2009: Implications of instrument design on measurement invariance*. Manuscript submitted for publication.

Wetzel, E., & Carstensen, C. H. (2013b). *Multidimensional modeling of response styles*. Manuscript submitted for publication.

Wetzel, E., & Carstensen, C. H. (2013c). *Reversed thresholds in the Partial Credit Model – A reason for collapsing categories?* Manuscript submitted for publication.

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant (Ed.), *The science of prevention. Methodological advances from alcohol and substance abuse research* (pp. 281–324). Washington, DC: American Psychological Association.

Wu, M. L. (2010). Measurement, sampling, and equating errors in large-scale assessments. *Educational Measurement: Issues and Practice*, *4*(29), 15–27.

Ziegler, M., Danay, E., Schölmerich, F., & Bühner, M. (2010). Predicting academic success with the Big 5 rated from different points of view: Self-rated, Other rated and faked. *European Journal of Personality*, *24*, 341–355. doi:10.1002/per.753

2. Appendix

2.1. Appendix A: Manuscript Wetzel, Böhnke, Carstensen, Ziegler, & Ostendorf (in press)

Do individual response styles matter? Assessing differential item functioning for men and women in the NEO-PI-R

Abstract

The occurrence of differential item functioning (DIF) for gender indicates that an instrument may not function equivalently for men and women. Aside from DIF effects, item responses in personality questionnaires can also be influenced by response styles. The aim of this study was to analyze the German NEO-PI-R regarding its differential item functioning for men and women while taking response styles into account. To this purpose, Mixed Rasch Models were estimated first to identify latent classes that differed in their response style. These latent classes were identified as extreme response style (ERS) and non-extreme response style (NERS). Then, DIF analyses were conducted separately for the different response styles and compared with DIF results for the complete sample. Several items especially on Neuroticism, Agreeableness, and Conscientiousness facets showed gender-DIF and thus function differentially between men and women. DIF results differed mainly in size between the complete sample and the response style subsamples, though DIF classification was overall consistent between ERS, NERS, and the complete sample.

Key words: differential item functioning, response styles, NEO-PI-R, Mixed Rasch Models

Introduction

Trait scores resulting from personality measures are often used to compare different groups of people, e.g., men and women. For example, using the revised NEO Personality Inventory (NEO-PI-R; Costa & McCrae, 1992), Costa, Terracciano, and McCrae (2001) found women to cross-culturally score higher on Neuroticism, Agreeableness, Warmth, and Openness to feelings, while men scored higher on Assertiveness and Openness to ideas. A meta-analysis on gender differences (Feingold, 1994) found men to report higher levels of Assertiveness and Self-esteem than women. Women were found to report higher levels of Extraversion, Anxiety, Trust, and Tender-mindedness. In addition to their application in research on gender differences, the application of personality measures is popular in personnel selection due to the value of personality traits in predicting job performance (Barrick & Mount, 1991). However, for comparisons of trait scores to be legitimate, whether for research or applied purposes, it is essential that the instrument functions equivalently across the groups of test-takers being compared. Otherwise, the differences between groups will be confounded with third variables. A potential confounding variable in personality questionnaires are response styles.

Differential Item Functioning

One method for examining whether the instrument functions equivalently is testing for differential item functioning (DIF). An item shows DIF if persons from different groups show different probabilities of responding correctly to the item although they have been matched on the underlying latent trait (Holland & Wainer, 1993). In the context of personality data, this corresponds to differences between groups (e.g., gender) in the probability of endorsing an item even if members of the different groups have the same latent trait level. In this case, the

choice of a response category is not only influenced by a person's trait level but instead additionally by a second dimension captured by the item. This second dimension can be another construct (e.g., another trait) or a different factor influencing the choice of a response category (e.g., response styles). If a second dimension influences responses, differences in test scores are due both to the trait under investigation and to the second dimension.

Testing for DIF is a standard tool in the area of ability testing, though applications to the personality domain are rare. One notable exception is a DIF analysis of the Abridged Big Five Circumplex (Mitchelson, Wicher, LeBreton, & Craig, 2009). Mitchelson et al. (2009) showed that 17 of the 45 scales contained gender-DIF. Another example are Smith and Reise (1998) who found gender-DIF on the Multidimensional Personality Questionnaire Stress Reaction Scale. Lastly, Reise, Smith, and Furr (2001) studied the measurement invariance of the NEO-PI-R's Neuroticism scale for men and women using DIF analyses. They found that several items showed DIF, some favoring men and others favoring women. Specifically, 16 items on the Neuroticism scale displayed DIF, six of which were located on the Anxiety facet.

Considering how widely used personality inventories are and since a lack of bias (e.g., in the form of DIF) is essential to fair testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999), it seems very important to broaden knowledge in the area of DIF in personality inventories. Furthermore, according to Wang (2008), when items contain DIF that is judged to be of practical significance, test scores are no longer comparable across groups. Consequently, the present study aimed at exploring DIF in one of the most widely used personality inventories, the NEO-PI-R, using a large sample.

Response Styles in Questionnaires

The discussion of DIF effects in questionnaires leaves a second, but equally important aspect, out of the picture: the question of the influence of individual response styles on test scores. In general, there are many potential factors that can influence the choice of a response category over and above a person's level on the trait being assessed, e.g., individual differences in response scale use, socially desirable responding, or differences in interpreting the item's content. Especially in personality inventories, which are usually administered using a polytomous response scale, participants often employ different response styles (consult Baumgartner & Steenkamp, 2001 for a review of common response styles). For example, in an analysis of the NEO-PI-R's short form, the NEO-FFI (Costa & McCrae, 1992), Austin, Deary, and Egan (2006) found that for Neuroticism, Agreeableness, and Conscientiousness, Mixed Rasch Models resulted in two-class solutions. Rost, Carstensen, and von Davier (1999) obtained the same result for Neuroticism, Agreeableness, Openness to experience, and Conscientiousness in the German NEO-FFI (Borkenau & Ostendorf, 1993). In both studies, participants in one class preferred extreme response categories (e.g., *strongly agree* or *strongly disagree*) while participants in the other class favored middle categories. The expressions used in this article for these two commonly occurring response styles are extreme response style (ERS; Baumgartner & Steenkamp, 2001) and non-extreme response style (NERS).

In line with these findings, Bolt and Johnson (2009) modeled a response style dimension in a multidimensional item response model which reduced the number of items showing DIF markedly compared to a unidimensional model that did not incorporate response styles. This indicates that differences in response style may have been responsible for the many DIF items found in the unidimensional model. Thus, response styles may have been the second dimension captured by the items in addition to the trait under investigation. The approach by Bolt and Johnson (2009) models response style as a continuous variable in contrast to Rost et

al. (1999) and Austin et al. (2006) who assume that response styles express qualitative differences. This paper takes the latter approach of understanding response styles as categorical latent variables where each person is assumed to show one of several distinct response styles. In both modeling alternatives it has been shown that individual differences in response styles can potentially distort the computation of sum scores and in consequence render comparisons of test takers based on their sum scores invalid. For example, if we have four items and a response vector (2, 2, 3, 3) for one person and (2, 2, 4, 4) for another person, based on sum scores it would be concluded that the second person has a higher trait level. This is not necessarily true if the second person has a preference for extreme categories while the first person has a preference for non-extreme categories; they might have the same level on the underlying latent trait. This illustrates why it is important to consider response styles.

Differences in response scale use can lead to different sum scores despite equal latent trait levels and are thus a potential source of DIF themselves. Therefore, this paper also addresses the question of whether differential item functioning varies with individual differences in response styles. In sum, the principal aim of this study was to examine two important sources of possible bias in personality test scores: We investigate whether gender-DIF exists in the NEO-PI-R while taking into account individual differences in response scale use. Thus, the objectives of this study were 1) to analyze the NEO-PI-R facets concerning gender-DIF and 2) to investigate whether DIF results differed between the complete sample and the separate response style subsamples.

Method

Instrument

The NEO-PI-R (Costa & McCrae, 1992) was applied in its German version (Ostendorf & Angleitner, 2004). The NEO-PI-R assesses the Big Five personality domains, namely Neuroticism (N), Extraversion (E), Openness to experience (O), Agreeableness (A), and Conscientiousness (C). Participants respond on a five-point scale with the response options *strongly disagree, disagree, neutral, agree, and strongly agree*. Rasch reliabilities for test scores on the 30 facets ranged from .55 for Openness to values to .84 for Depression.

Samples

The participants consisted of the non-clinical standardization sample for the German NEO-PI-R (Ostendorf & Angleitner, 2004). In total, 11,724 persons were part of the standardization sample. For this study, the sample was randomly divided into two halves, the first half being used for the following analyses ("analyses sample") and the second half being reserved for validation purposes (see Wetzel, Carstensen, and Böhnke, in press, for a summary of the validation results concerning response styles in the NEO-PI-R). Thus, the analyses sample used in the following contained 5,862 participants (63.8% women and 36.2% men) who were between 16 and 87 years old ($M = 29.9$, $SD = 12.2$). Cohen's d for gender differences (Cohen, 1988) on the 30 facets ranged from -0.02 (Openness to values) to -0.54 (Anxiety). Medium to large gender differences in scores occurred on some facets of all Big Five domains, most notably Neuroticism for which the sum score showed a d of -0.48. Gender differences in our sample are consistent with those found in previous studies (e.g., Feingold, 1994; Costa et al., 2001).

Analyses

The data were analyzed in two steps. First, in the pre-analyses, Mixed Rasch Models were applied to test the homogeneity of the participants' response behavior. Second, analyses of differential item functioning regarding gender were performed for the complete sample and the different response style subsamples.

Pre-analyses: Response Styles in the NEO-PI-R

Mixed Rasch Models (MRMs; Rost, 1990; Rost, 1991) were conducted first to test whether latent classes of respondents existed who differed systematically in their response behavior. MRMs combine Rasch analysis with latent class analysis (LCA). In an LCA, participants are allocated to qualitatively different latent classes; e.g., latent classes differing in their response style. An MRM then allows that the Rasch Model (RM) holds within each latent class but with different parameters between the latent classes. Thus, both qualitative (LCA) and quantitative (RM) differences between participants are captured in Mixed Rasch Models. To model the response probabilities for ordered polytomous response categories the Mixed Rasch Model based on Masters' (1982) partial credit model (PCM) developed by Rost (1991) is employed. This model specifies threshold parameters τ_{isg} (s = 0, …, m) which model the distribution of the responses on item i over a response scale with $m+1$ categories for each class g. The mean of these threshold parameters can be parameterized as δ_{ig} and can be interpreted as the mean item difficulty. Different response styles are identified if MRM classes are found that show both of the following two properties:

1) Differential use of the response scale between classes. For instance, in one class respondents tend to endorse the more extreme response categories more frequently whereas respondents from another class tend to use the middle categories more

frequently. This is expressed through different sets of threshold parameters τ_{isg} for the different classes g.

2) The latent trait is the same in all classes. In a Rasch Model, this is true if the item location parameters for each item δ_{ig} are the same over latent classes.

If in all classes derived from an MRM analysis the same trait is measured (property 2), differences between classes can only be due to differences in response style (property 1) and DIF with respect to one underlying trait can be analyzed in different response styles. To this end, two mixed partial credit models (mixed PCMs) were compared regarding model fit: a regular mixed PCM in which all threshold and location parameters were estimated freely and a constrained mixed PCM in which item location parameters δ_{ig} were constrained to be equal between the latent classes, ensuring the same construct being measured in all latent classes. In analogy to the model comparison introduced by Rost and von Davier (1995), we compared a (freely estimated) two-class mixed PCM to a model with a single latent trait[2]. Hence, if the constrained mixed PCM shows a better fit, the same dimension is being measured in both classes and the heterogeneity between classes is only due to differences in response scale use. This comparison also tests the assumption of person homogeneity (with respect to item location parameters) in terms of item response model fit testing. All models were estimated using WINMIRA (von Davier, 2001).

Model fit was assessed using the Consistent Akaike's Information Criterion (CAIC; Bozdogan, 1987). The better-fitting model is the one with a lower CAIC value. The score distributions were approximated using a logistic distribution in WINMIRA if the approximation was sufficiently close to the unconstrained score distribution (indicated by an RMSEA below .08). The mixed PCM and constrained mixed PCM were estimated separately for each

[2] Rost and von Davier (1995) used a one-class model instead.

facet of the NEO-PI-R. Drawing upon the results from other studies applying MRMs to a NEO questionnaire (Rost et al., 1999; Austin et al., 2006), one to four-class solutions were estimated. Models were fitted with 10 random starting values to avoid finding local maxima instead of the global maximum likelihood solution (Rost, 1991).

Facets for which the constrained mixed PCM showed a better fit (homogeneity, see above) were included in the subsequent separate DIF analyses by response style since it could be concluded that the same trait was being measured in the latent classes. On the other hand, facets for which the mixed PCM showed a better fit were excluded from the separate DIF analyses by response style because different traits appeared to be measured in the latent classes, rendering them incomparable. For the facets remaining after the model comparison, classes were interpreted as distinct response styles using item threshold parameters. For each of these facets, the probability of belonging to each of the different latent classes (e.g., ERS and NERS) was computed in the constrained mixed PCM for every participant. Participants were allocated to the latent class for which their probability of class membership was highest. The resulting class membership information was used to examine whether a relationship existed between class membership and gender by computing χ^2-tests.

Gender-DIF in the NEO-PI-R for the Complete Sample and Different Response Styles

Analyses of differential item functioning (DIF) for gender were conducted using a DIF estimate in the framework of Item Response Theory as implemented in ConQuest (Wu, Adams, Wilson, & Haldane, 2007). In ConQuest, a model was specified to include the interaction between item difficulty and a grouping variable for gender. The interaction term implies that item difficulty parameters are estimated separately for men and women and calibrated onto the same scale. The DIF estimates are obtained by computing the difference between the two item difficulty parameters for each item (Le, 2009). These DIF estimates can be tested for

significance using a χ^2-test. Since the statistical significance of the DIF estimate depends on sample size it is advisable to evaluate the DIF's absolute size additionally. To this purpose, a classification system has been developed by Educational Testing Service (ETS; Zieky, 1993) which categorizes items as containing negligible, slight to moderate, or moderate to large DIF. We applied this classification system, though logits were used as the measurement unit instead of Delta units. According to this classification, a DIF contrast of less than .25 logits is negligible (category A), a DIF contrast between .25 and .37 logits is slight to moderate (category B), and a DIF contrast equal to or above .38 is moderate to large (category C; Zieky, 1993). In addition, the χ^2-test of the DIF estimate had to be significant at the .01 level for an item to be classified as a B or C item (Le, 2009). Thus, for each item we can conclude whether men and women differ in their endorsement probability of this item's categories despite having the same latent trait measure. DIF analyses were first performed separately for each of the facets on the complete sample. Then, DIF analyses were repeated for the facets that showed different (but trait-homogeneous) response styles. DIF results were compared between the complete sample and the response style subsamples.

Results

In the following, results will be reported first for the pre-analyses concerning the identification of response styles. Second, DIF results will be described for the complete sample and the different response style subsamples and comparisons will be drawn between these samples.

Pre-analyses: Response Styles in the NEO-PI-R

The mixed PCM and constrained mixed PCM were estimated for each of the 30 facets of the NEO-PI-R for the analyses sample. Model fit was compared using the CAIC (Bozdogan, 1987). The log-likelihood and CAIC values for the two models are depicted in Table 1. Five facets could not be included in the model comparison due to the occurrence of categories with response frequencies of zero (null categories). Null categories cause estimation problems in the mixed PCM as the response probabilities for these categories approach zero. Regarding the remaining 25 facets, the constrained mixed PCM showed a better fit for 16 facets while the mixed PCM showed a better fit for ten facets (Table 1). For these 16 facets the better-fitting constrained mixed PCM also indicated that the central IRT assumption of person homogeneity was given. For 14 of these 16 facets a two-class solution was appropriate (and better-fitting than the one class solution) since the two classes could be interpreted as distinct response styles unambiguously. Three- and four-class solutions in most cases showed a better fit than two-class solutions as indicated by the CAIC. However, except for Openness to actions and Deliberation (see below), the latent classes could not be interpreted clearly as distinct response styles. Often, the third and/or fourth class was similar to the first or second class and did not appear to be a qualitatively different response style. Instead, these classes mainly differed from the first two classes regarding their item parameters for a few items, though the general pattern was very similar to the first two classes. Furthermore, the allocation of participants to the third and fourth classes involved a high degree of uncertainty. Thus, we decided to interpret the more parsimonious two-class solutions (see Rost et al., 1999 for a similar situation and reasoning regarding a mixed Rasch analysis of the NEO-FFI).

Table 1

Comparison of Model Fit for the Mixed Partial Credit Model (mixed PCM) and the Constrained Mixed PCM for the NEO-PI-R

Facet		Mixed PCM LL	Mixed PCM CAIC	Constrained mixed PCM LL	Constrained mixed PCM CAIC	N
Neuroticism						
N1 Anxiety		-57914.04	116475.55	-57937.38	**116444.92**	5789
N2 Angry hostility		-58524.91	117697.04	-58458.95	**117487.84**	5768
N3 Depression		-57060.57	**114768.79**	-57114.06	114798.43	5804
N4 Self-consciousness*		-60661.81	**122532.17**	-60709.60	122550.39	5816
N5 Impulsiveness		-60791.16	122229.78	-60829.54	**122229.25**	5789
N6 Vulnerability*		-52787.58	**106783.03**	-52854.33	106839.24	5785
Extraversion						
E1 Warmth		-51568.99	**103785.36**	-51664.50	103899.09	5782
E2 Gregariousness		-58637.78	117922.52	-58672.09	**117913.90**	5746
E3 Assertiveness		-57565.96	115779.34	-57603.21	**115776.54**	5785
E4 Activity		-58098.99	116845.55	-57873.30	**116316.85**	5797
E5 Excitement-seeking*		-66915.58	135039.28	-66943.99	**135018.78**	5796
E6 Positive emotions*		-54077.43	109362.84	nc	nc	5790
Openness to experience						
O1 Fantasy		-56058.18	**112763.46**	-56217.69	113005.22	5758
O2 Aesthetics*		-56524.36	**114256.96**	-56611.67	114354.27	5802
O3 Feelings		-50433.73	101515.02	-50460.44	**101491.12**	5797
O4 Actions*	2 classes	-59671.26	120551.17	-59690.41	120512.11	5821
	3 classes	-59373.21	120554.56	-59393.49	**120440.43**	5821
O5 Ideas*		-58149.13	**117506.70**	-58340.12	117811.35	5811
O6 Values		-58064.11	116775.60	-58075.17	**116720.44**	5782
Agreeableness						
A1 Trust		nc	nc	-55921.67	112413.47	5786
A2 Straightforwardness		-59550.77	**119749.15**	-59904.27	120378.83	5801
A3 Altruism*		nc	nc	nc	nc	5790
A4 Compliance*		-58627.70	118463.69	-58635.43	**118401.82**	5804
A5 Modesty		nc	nc	-58435.78	117441.59	5777
A6 Tender-mindedness		-54184.46	109016.41	-54203.30	**108976.77**	5790
Conscientiousness						
C1 Competence		-52383.08	105413.52	nc	nc	5779
C2 Order*		-59231.31	119671.13	-59268.22	**119667.59**	5814
C3 Dutifulness		-53637.04	**107921.57**	-53712.21	107994.60	5791
C4 Achievement striving*		-59358.01	119924.31	-59381.33	**119893.61**	5804
C5 Self-discipline		-55376.84	111401.10	-55390.49	**111351.11**	5786
C6 Delibera-	2 classes	-56722.95	114654.30	-56736.63	114604.31	5809
tion*	3 classes	-56124.80	114057.35	-56186.08	**114025.25**	5809

Note. Number of parameters for mixed PCM = 67, number of parameters for constrained mixed PCM = 59, LL = log-likelihood, CAIC = Consistent Akaike's Information Criterion, nc = null categories. The CAIC of the better-fitting model is depicted in boldface.

* Facets for which the score distribution in WINMIRA was not approximated. Number of parameters for mixed PCM = 125, number of parameters for constrained mixed PCM = 117.

For facets in which the constrained mixed PCM was the better-fitting model, threshold plots, which show the threshold parameters for each item on the respective facet, were inspected to interpret classes with regard to the class members' response behavior. Threshold parameters revealed two consistently occurring response styles, namely extreme response style (ERS) and non-extreme response style (NERS). Examples for the two response styles are depicted in Figure 1 for the facet Achievement striving. NERS is characterized by widely spaced first and fourth thresholds while the second and third thresholds are close together. In contrast, with ERS all thresholds are nearby each other, sometimes even overlapping. In the context of endorsing the NEO-PI-R's items it follows that for participants with an extreme response style, the increase in trait level necessary to endorse the fifth response category instead of the first is very small. For NERS however, both the trait levels needed to endorse the lowest or the highest response category are at the extreme ends of the trait continuum while the second to fourth response categories are endorsed at average trait levels. Thus, ERS is characterized by a strong preference for extreme categories (*strongly disagree* and *strongly agree*) and NERS is overall characterized by a preference for middle categories (*disagree* and *agree*). However, this does not mean that persons allocated to the ERS or NERS group never chose a middle or extreme category, respectively. Instead, the distribution of the category frequencies differs between NERS and ERS. The categories *disagree, neutral,* and *agree* show higher category frequencies for NERS, while *strongly disagree* and *strongly agree* show higher category frequencies for ERS (for an illustration see Figure 2). Note that respondents who used the response scale evenly are more similar to NERS than to ERS and thus have a higher probability of being allocated to this class which also contains the majority of the participants. Category frequencies show that the extreme categories were also endorsed in the

NERS class (Figure 2). Since Mixed Rasch Models were computed separately for each facet, participants were allocated to the response styles anew for each facet, so it is possible that a person for example belonged to the ERS group for one facet and to the NERS group for another facet. Nevertheless, Wetzel et al. (in press) showed that for the majority of the participants in this sample (ca. 80%) the response style was used consistently across facets. Thus, most participants used the same response style independently of the trait that was being assessed. For each facet, the probability of belonging to NERS or ERS was computed for every participant and participants were allocated to the class with the higher probability. Over all facets and domains, the maximum probability for class assignment ranged from .60 to 1.00 (M = .86, SD = .05), indicating that participants could be assigned to the NERS or ERS group with high certainty. Pertaining class size, the class of participants using NERS was consistently larger compared to the class using ERS. Between sixty-six and 76% of the respondents were allocated to NERS while 23% to 33% were allocated to ERS.

For Openness to actions and Deliberation, three classes were found that differed with respect to their response behavior and were homogeneous regarding the trait being assessed (as indicated by the better fit of the constrained mixed PCM). For these two facets, non-extreme responders were allocated to two classes that differed in their use of the middle category. The first class of NERS appeared to use the middle category rarely (nearly overlapping second and third threshold parameters) whereas the second class of NERS did not employ the middle category at all (second and third threshold parameters reversed and widely spaced).

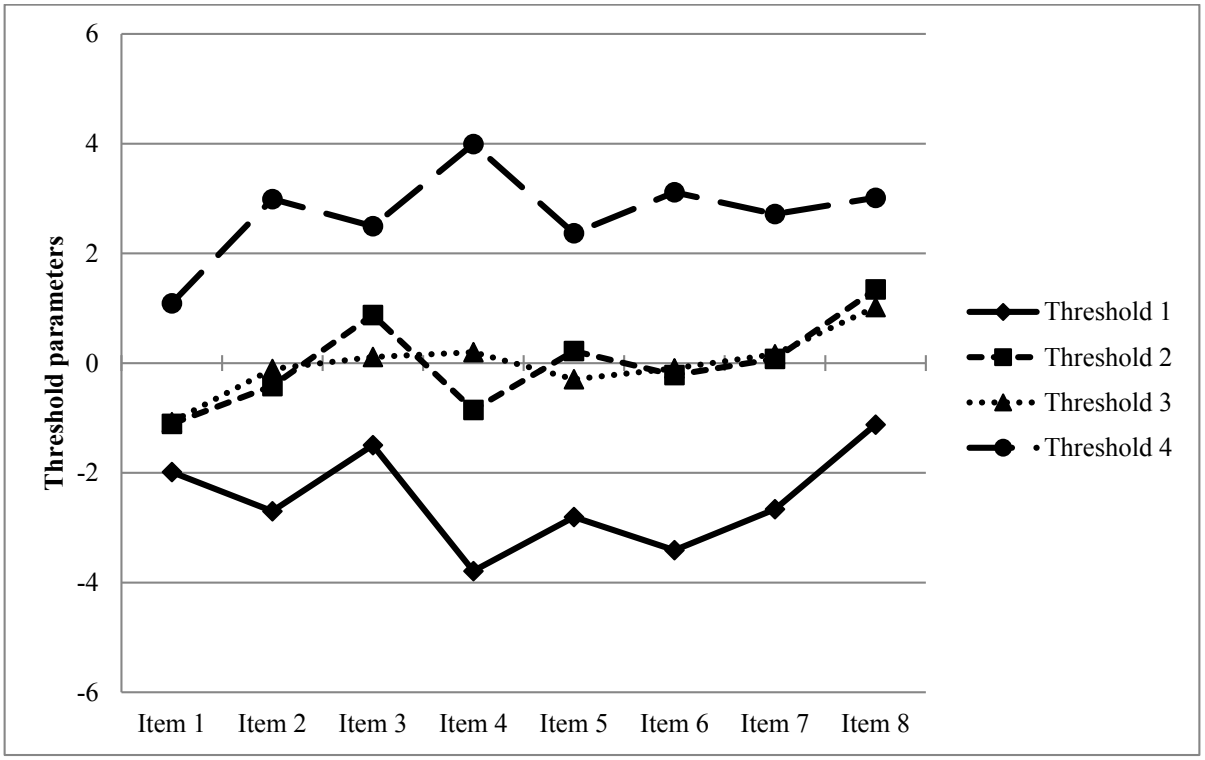

Figure 1.a. Non-extreme response style on the facet Achievement striving.

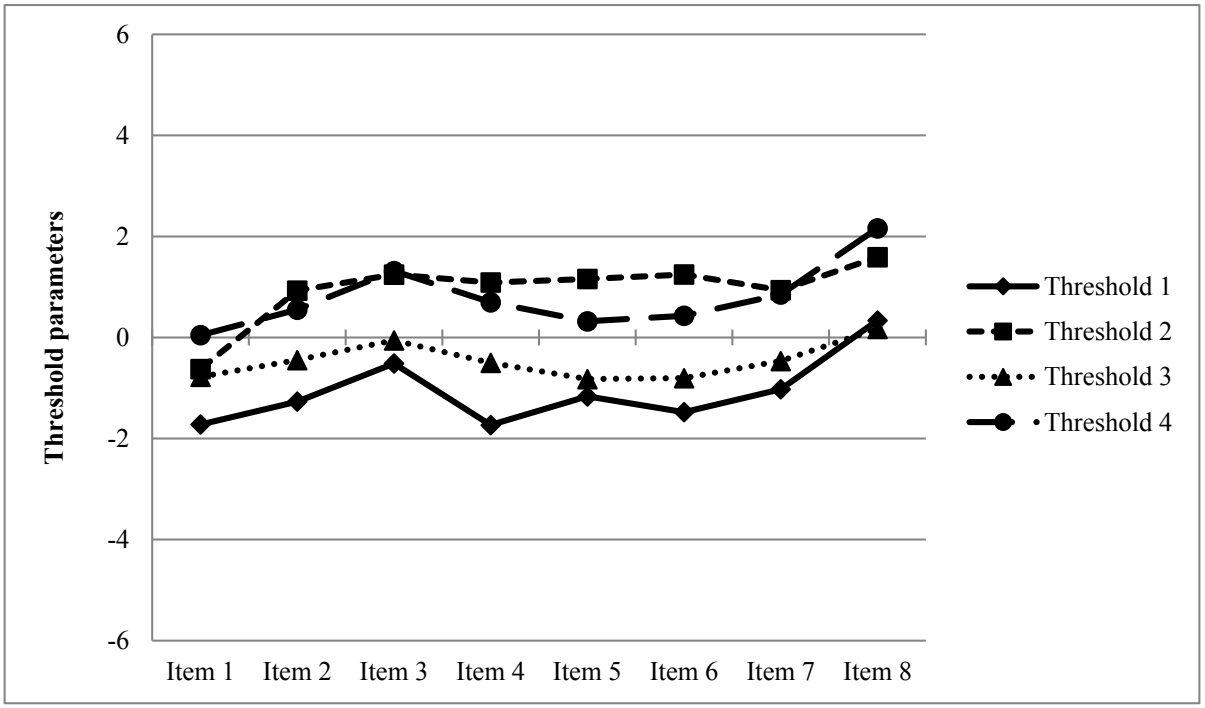

Figure 1.b. Extreme response style on the facet Achievement striving.

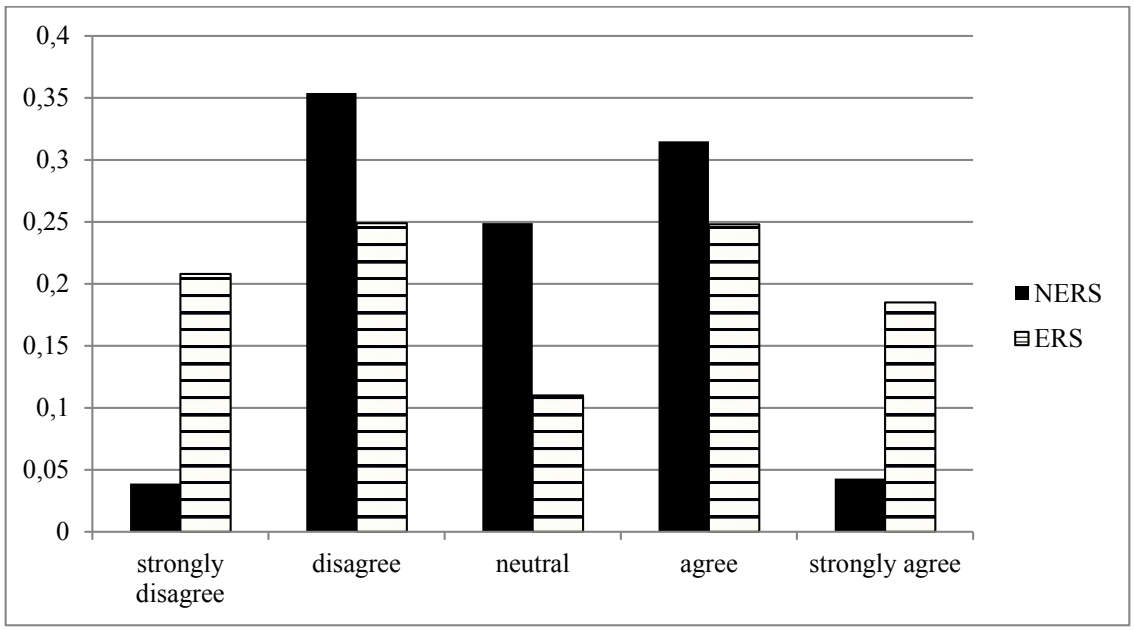

Figure 2. Category frequencies for non-extreme response style (NERS) and extreme response style (ERS) on item 31 (facet Anxiety).

χ^2-tests were computed to examine whether a relationship between gender and response style existed. As depicted in Table 2, this was not the case for five facets. For eleven facets, the χ^2-test yielded a significant result implying that for these facets gender and response styles were related. However, effect sizes ω for these relationships can be classified as small according to Cohen (1988).

Gender-DIF in the NEO-PI-R for the Complete Sample and the Different Response Styles

DIF analyses were first conducted separately for all of the NEO-PI-R's facets using the complete sample. Then, separate DIF analyses were conducted within the response style classes for each of the 16 facets in which trait-homogenous classes were found and in which threshold parameters could be clearly interpreted as distinct response styles. The complete sample contained the 5,862 participants in the analyses sample.

Table 2

Results on the Relationship between Gender and Class Membership

Facet	χ^2	df	p	ω
Neuroticism				
N1 Anxiety	1.44	1	.232	.02
N2 Angry hostility	18.77	1	<.001	.06
N5 Impulsiveness	6.00	1	.015	.03
Extraversion				
E2 Gregariousness	2.07	1	.152	.02
E3 Assertiveness	4.01	1	.047	.03
E4 Activity	4.26	1	.041	.03
E5 Excitement-seeking	3.00	1	.086	.02
Openness				
O3 Feelings	57.11	1	<.001	.10
O4 Actions	6.43	2	.040	.03
O6 Values	35.83	1	<.001	.08
Agreeableness				
A4 Compliance	1.40	1	.239	.02
A6 Tender-mindedness	13.81	1	<.001	.05
Conscientiousness				
C2 Order	0.35	1	.560	.01
C4 Achievement striving	7.94	1	.005	.04
C5 Self-discipline	4.33	1	.039	.03
C6 Deliberation	18.92	2	<.001	.06

Note. Only facets in which the constrained mixed Partial Credit Model showed a better fit were used for this analysis.

Gender-DIF was found on several items in the NEO-PI-R (Table 3). Regarding all of the 30 NEO-PI-R facets, 24 items contained slight to moderate and seven items contained moderate to large gender-DIF in the complete sample. Positive DIF values indicate DIF favoring men, which in the present context can be interpreted as items on which men have a higher probability of endorsing a higher response category compared to women of the same trait level. DIF in both directions occurred. The direction of DIF in the complete sample was almost balanced with 17 items favoring men and 14 items favoring women. Overall, DIF estimates ranged from 0 (item 6 on E5 and item 6 on C4) to .52 logits (item 5 on N2 and item 6 on O2). The χ^2-test of the DIF estimate was significant at the .01 level for all B and C DIF category items in Table 3. Concerning Neuroticism, three items (item 2 on N1, item 5 and item 6 on

N2) displayed moderate to large DIF. Extraversion was the Big Five domain with the least DIF items. Two items (item 4 on E5 and item 4 on E6) showed slight to moderate DIF. Regarding Openness, the facet Openness to aesthetics contained four DIF items. The facet Modesty (on Agreeableness) was the one with the most DIF items (5) of all facets, though all DIF items were in the slight to moderate category. Conscientiousness contained two large DIF items (item 1 on C4 and item 1 on C6). The separate DIF analyses for the response style subsamples yielded 12 items with DIF contrasts ≥ .25 and six items with DIF contrasts ≥ .38 in the NERS group. Regarding the ERS group, seven items showed slight to moderate DIF and three items showed moderate to large DIF.

In comparing DIF results between the complete sample and the separate response styles, two points are worth noting. First, the items on which DIF occurred were very consistent between the complete sample and the two response styles. There are only few exceptions, where an item showed DIF in one or two of the subsamples but not in the other(s). E.g., item 3 on Openness to feelings showed gender-DIF for the NERS group while it was inconspicuous for the ERS group and the complete sample. Second, on items that showed DIF in more than one response style and/or the complete sample, DIF values differed to varying degrees between the samples. For some items (e.g., item 5 on facet N2), DIF values were practically identical (.52 for the complete sample, .50 for NERS, and .52 for ERS). For other items, DIF values differed strongly, leading to incongruent classifications as A, B, or C items (e.g., item 2 on facet N5). In summary, over all common items between the samples, differences in classification mainly pertain to neighboring categories (e.g., item 6 on N2 contains moderate to large DIF in the complete sample and the NERS subsample but slight to moderate DIF in the ERS subsample).

Table 3

Results from Differential Item Functioning (DIF) Analyses in ConQuest

	DIF Estimate (SE)				Item content
Facet	**Complete sample**	**NERS**	**ERS**	**NERS 2**	
Neuroticism					
N1 Anxiety					
Item 2	**-.38 (.02)**	**-.50 (.03)**	*-.24 (.03)*	n/a	I am easily frightened.
Item 6	.28 (.02)	.36 (.03)	*.17 (.03)*	n/a	I am often worried about things that might go wrong.
N2 Angry hostility					
Item 5	**.52 (.02)**	**.50 (.03)**	**.52 (.04)**	n/a	I often don't like the people I have to deal with.
Item 6	**-.44 (.02)**	**-.48 (.03)**	-.35 (.03)	n/a	I don't lose my cool easily.
N3 Depression					
Item 4	.28 (.02)	n/a	n/a	n/a	Sometimes I have a strong feeling of guilt and sinfulness.
N4 Self-consciousness					
Item 4	-.37 (.02)	n/a	n/a	n/a	I'm not easily embarrassed when others mock or ridicule me.
Item 6	.29 (.02)	n/a	n/a	n/a	I feel comfortable in the presence of my boss or other authorities.
N5 Impulsiveness					
Item 2	.32 (.02)	**.40 (.03)**	*.20 (.03)*	n/a	It's hard for me to resist my desires.
Item 8	-.27 (.05)	-.28 (.07)	-.28 (.09)	n/a	I am always capable of keeping my feelings under control.
N6 Vulnerability	A	n/a	n/a	n/a	
Extraversion					
E1 Warmth	A	n/a	n/a	n/a	
E2 Gregariousness	A	A	A	n/a	
E3 Assertiveness	A	A	A	n/a	
E4 Activity	A	A	A	n/a	
E5 Excitement-seeking					
Item 4	.25 (.02)	*.24 (.02)*	.26 (.02)	n/a	Whenever possible, I avoid watching shocking or scary movies.

(continued)

	DIF Estimate (SE)				Item content
Facet	**Complete sample**	**NERS**	**ERS**	**NERS 2**	
E6 Positive emotions					
Item 4	-.28 (.02)	n/a	n/a	n/a	Sometimes I overflow with happiness.
Openness					
O1 Fantasy	A	n/a	n/a	n/a	
O2 Aesthetics					
Item 2	.35 (.02)	n/a	n/a	n/a	Sometimes I get completely absorbed in music I listen to.
Item 3	**-.48 (.02)**	n/a	n/a	n/a	It bores me to watch ballet or modern dance.
Item 6	**.52 (.02)**	n/a	n/a	n/a	Certain types of music fascinate me endlessly.
Item 8	-.32 (.06)	n/a	n/a	n/a	I prefer literature that emphasizes emotions and fantasy more than action plots.
O3 Feelings					
Item 3	*-.23 (.02)*	-.26 (.03)	*-.16 (.04)*	n/a	My feelings toward things are important to me.
O4 Actions					
Item 3	*.18 (.02)*	.26 (.04)	*.09 (.04)*	*.17 (.04)*	Once I found a way to do something, I stick to it.
Item 6	-.34 (.02)	-.35 (.04)	*-.24 (.03)*	**-.41 (.03)**	Sometimes I only make changes in my home to try out something new.
Item 8	.25 (.05)	*.24 (.10)*	*.18 (.09)*	.32 (.09)	When I go somewhere, I always take the proven route.
O5 Ideas	A	n/a	n/a	n/a	
O6 Values	A	A	A	n/a	
Agreeableness					
A1 Trust	A	n/a	n/a	n/a	
A2 Straight-forwardness					
Item 8	.34 (.06)	n/a	n/a	n/a	I am proud of my skillfulness in dealing with others.
A3 Altruism	A	n/a	n/a	n/a	
A4 Compliance					
Item 1	*-.22 (.02)*	*-.20 (.03)*	-.26 (.04)	n/a	I prefer to cooperate with others than to compete with them.
Item 2	-.34 (.02)	**-.**37 (.03)	-.28 (.03)	n/a	When necessary, I can be sarcastic and mocking.

(continued)

	DIF Estimate (SE)				Item content
Facet	**Complete sample**	**NERS**	**ERS**	**NERS 2**	
A5 Modesty					
Item 2	.27 (.02)	n/a	n/a	n/a	I'd rather not talk about myself and my accomplishments.
Item 3	-.28 (.02)	n/a	n/a	n/a	I am better than most people and I know it.
Item 4	.30 (.02)	n/a	n/a	n/a	I try to be humble.
Item 7	.29 (.02)	n/a	n/a	n/a	I prefer praising others to being praised.
Item 8	-.32 (.06)	n/a	n/a	n/a	I believe that I am superior to others.
A6 Tender-mindedness					
Item 8	.31 (.06)	.33 (.08)	*.21 (.10)*	n/a	I'd rather be known for being kind than for being just.
Conscientiousness					
C1 Competence					
Item 3	.34 (.03)	n/a	n/a	n/a	I keep myself informed and usually make intelligent decisions.
C2 Order					
Item 2	-.28 (.02)	-.31 (.03)	-.27 (.03)	n/a	I keep my things tidy and clean.
Item 3	.27 (.02)	.31 (.03)	*.21 (.03)*	n/a	I do not proceed in a very systematic manner.
C3 Dutifulness	A	n/a	n/a	n/a	
C4 Achievement striving					
Item 1	**-.42 (.02)**	**-.42 (.03)**	**-.41 (.04)**	n/a	I am easygoing and unconcerned.
C5 Self-discipline					
Item 1	*-.20 (.02)*	-.26 (.03)	*-.12 (.03)*	n/a	I'm good at making a schedule so I get everything done on time.
Item 8	.31 (.06)	.31 (.08)	.34 (.09)	n/a	I possess a high degree of self-discipline.
C6 Deliberation					
Item 1	**-.45 (.02)**	**-.44 (.04)**	**-.49 (.04)**	-.29 (.03)	I've done some stupid things in my life.
Item 4	*.23 (.02)*	.25 (.04)	*.21 (.04)*	.28 (.04)	Before I act, I generally think through the possible consequences.

Note. DIF values < .25 (A items) are in italics, DIF values ≥ .25 (B items) are depicted in regular font, DIF values ≥ .38 (C items) are in boldface. A = only A items were found on the respective facet, n/a = item or facet was not analyzed separately by response style. Positive values

indicate DIF favoring men, negative values indicate DIF favoring women. Item content consists of the German NEO-PI-R items translated into English by the authors.

Discussion

Differential item functioning poses a serious threat to researchers and practitioners using psychological tests. Since personality questionnaires are increasingly popular in a variety of assessment contexts, systematic research regarding DIF in this area is needed. While some prior studies exist, individual response styles, which might distort DIF analyses, have not been considered yet. The present study tried to elucidate the amount of DIF in a large population sample and compared whether the results differed between response styles. The results showed that for most of the facets analyzed regarding response styles, two different response styles occurred, i.e., non-extreme response style and extreme response style. DIF analyses revealed the presence of items that did not function equivalently between men and women on many facets. It was found that controlling for response style had little effect on most DIF items. However, some items were only identified as containing gender-DIF when response style was controlled for. Finally, the amount of DIF for some items changed when response style was controlled for.

Gender-DIF in the NEO-PI-R

In this study, DIF was analyzed regarding gender. Other groups are conceivable in which DIF might exist, for example different age groups, different ethnic groups, or different cultural contexts. Overall, DIF results were consistent between the complete sample, NERS, and ERS with respect to the number of DIF items on the common facets and the direction of DIF. A possible reason for this finding is that in our sample, effect sizes for the relationship between response styles and gender were small. If a stronger relationship between response styles and the group variable of interest existed, it would be more likely for DIF results to be affected (for an example see Bolt & Johnson, 2009). However, it was also found that for some items the status of being classified as containing DIF or not changed when response style was

controlled for. Even though this occurred only for few items, this finding, if replicated, shows the potential importance of controlling for response styles in the analysis of DIF. The influence of response styles on DIF results is further indicated by differences in DIF size between the samples for several items. In total, though, the classification of DIF as negligible, slight to moderate, or moderate to large was overall consistent between the samples, indicating that the practical implications of these differences may be minor, especially since only few items demonstrated moderate to large DIF. It follows that, overall, the results in our study indicate that gender-DIF and response styles can be interpreted as two rather independent influences on item responses and thus, differences in response styles are only one of many possible reasons gender-DIF can be attributed to. Nevertheless, one has to keep in mind that DIF is of interest with respect to other individual differences variables as well, e.g., age.

Our DIF results concerning Neuroticism differed from Reise et al. (2001) in that their study found more DIF items. In total, Reise et al. found 16 DIF items while we found nine DIF items on the Neuroticism facets. For example, concerning Vulnerability, three items in Reise et al. demonstrated DIF while none did in our study. Six of the DIF items overlap between the two studies, e.g., in both studies items five and six on Angry hostility displayed DIF. These differences are possibly due to the different IRT models (mixed PCM and generalized PCM), the different samples (Germans and Americans), and the different versions of the instrument (German and English NEO-PI-R). Furthermore, we evaluated DIF items according to the ETS classification system, taking into account size as well as significance of DIF while Reise et al. considered all items with significant χ^2 values as displaying DIF. Our comparison between the constrained mixed PCM and the mixed PCM showed that some of the facets were not unidimensional. Since the dimensionality of the facets was not taken into account in Reise

et al.'s study, more items would be expected to show DIF in their analyses due to multidimensionality. On items that showed DIF in both studies, the direction of DIF was the same across both studies.

Following the ETS procedure to interpreting DIF data (Zieky, 1993) both B and C items should be reviewed by expert teams to examine possible causes for the occurrence of DIF and to decide whether the items can be considered fair and suitable for further use. Thus, in a revision of the NEO-PI-R a review of items showing gender-DIF of slight to moderate and moderate to large size appears appropriate in order to ensure the fairness of all items to both gender groups. Considering our results this would mainly pertain to items on several Neuroticism, Agreeableness, and Conscientiousness facets. Especially the facets Anxiety, Angry hostility, Achievement striving, and Deliberation contain items displaying large DIF.

Response Styles in the NEO-PI-R

In other studies investigating response styles using MRMs (e.g., Austin et al., 2006), a two-class solution consisting of one class of extreme responders and one class of non-extreme responders usually fit the data best. In our study, two classes were adequate to describe the participants' response behavior on most facets as well. However, for about half of the NEO-PI-R's facets, the latent classes derived from the MRMs showed trait-heterogeneity. Thus, these facets do not appear to be unidimensional. For Openness to actions and Deliberation, a third class emerged in addition to NERS and ERS. This class was also characterized by non-extreme response style but differed from the other NERS class in its use of the middle category. No consistent relationship between gender and class membership (to a certain response style) could be established. This coincides with the inconclusive nature of previous results (e.g., Austin et al., 2006).

In this paper, response style was viewed as a discrete variable. That is, it was assumed that participants differ qualitatively in their response behavior, allowing them to be allocated to different categories of response styles. Our results suggest that response styles can be modeled using two classes, namely NERS and ERS. In future research, these results could be contrasted with the modeling of response styles as a continuous variable (e.g., Bolt & Johnson, 2009). Our approach to modeling response styles applied a model fit comparison between a mixed partial credit model and a constrained mixed partial credit model. This approach is similar to Rost et al. (1999) and Austin et al. (2006) who also computed Mixed Rasch Models but goes beyond their work in ensuring that the latent classes derived from the MRMs differ only in their response style but not in other aspects such as the trait that is being assessed.

Implications of DIF and Response Styles

The occurrence of both response styles and DIF can be viewed as interfering with the measurement of the intended trait. DIF can raise or lower sum scores, depending on its direction. DIF favoring women might raise women's scores relative to men's scores while DIF favoring men might do the opposite (Reise et al., 2001). It follows that DIF may produce a bias in the computation of total scores. Using gender-specific norms (as provided in the NEO-PI-R) may reduce the influence of item-level DIF on the sum score level. The occurrence of DIF raises some doubts with regard to the use of personality questionnaires where gender is an issue since comparisons between trait scores might be rendered invalid by DIF (Wang, 2008). Thus, studies investigating gender differences or practitioners using such a questionnaire for selection purposes should consider removing critical items. Otherwise, results might be artificially distorted or adverse impact might occur. A limitation that has to be noted is that gender concepts might differ between different cultures. Therefore, the findings of this study

might not generalize across different cultures. However, it also has to be noted that the Five Factor Model has been replicated across several languages and cultures (Costa et al., 2001).

Similarly, response styles can lead to biased sum scores. For example, two persons may feel equally strong about the statement made in an item but then end up choosing different response categories (e.g., *agree* and *strongly agree*) because one has a preference for extreme responding and the other has a preference for mid-scale responding. According to Austin et al. (2006), by tendency, extreme responders are assigned more extreme trait scores than non-extreme responders of similar trait levels. The assumption underlying comparisons of sum scores is that all test takers use the response scale in the same manner. When participants differ in their response styles, inferences drawn from the comparison of sum scores may be invalid. Furthermore, since response styles are consistent across traits for the majority of the respondents (Wetzel et al., in press), response styles appear to be a systematic influence on item responses. Thus, further research should explore ways of lessening the impact of response styles by taking individual differences in response styles into account when evaluating and interpreting a person's test scores.

Considering ways to lessen the impact of response styles could also focus on the questionnaire itself. For example, a dichotomous rating format (provided it is appropriate to the questionnaire's contents) precludes preferring the middle category or extreme categories. Even-numbered answer formats might have the same effect. Moreover, there is a recent development showing that forced-choice item formats can yield normative data (Brown & Maydeu-Olivares, 2011). Reverse-coded items as they are present in the NEO-PI-R may be useful in dealing with response styles such as acquiescence (however, see Marsh, 1996) but they do not counteract NERS or ERS. Modeling response style as a unique dimension in a multidimensional model as applied in Bolt and Johnson (2009) appears to be an interesting alternative

to the categorical approach taken here. Both approaches should be compared with respect to their ability of partialing out the response styles' influence on trait scores.

Conclusion

Gender-DIF occurred on several Neuroticism, Agreeableness, and Conscientiousness items on the German NEO-PI-R, while Extraversion and Openness to experience appear to overall function equivalently for men and women. Response styles characterized by extreme responding and non-extreme responding were identified on many NEO-PI-R facets. Both response styles and DIF can potentially interfere with trait measurement if they are not considered. DIF results showed differences especially in size between the complete sample, non-extreme responders, and extreme responders, though the classification of DIF size was overall consistent between the samples for most items. The overall consistency of gender-DIF results across response style subsamples indicates that gender-DIF and response styles appear to independently influence item responses. Our findings show that it is important to consider both response styles and DIF in analyzing personality data.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Education Research Association.

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

Barrick, M. R., & Mount, M. K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology*, *44*(1), 1–26. doi:10.1111/j.1744-6570.1991.tb00688.x

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

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. doi:10.1177/0146621608329891

Borkenau, P., & Ostendorf, F. (1993). *NEO-Fünf-Faktoren-Inventar (NEO-FFI) nach Costa und McCrae*. Göttingen, Germany: Hogrefe.

Bozdogan, H. (1987). Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, *52*(3), 345–370. doi:10.1007/BF02294361

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. New York, NY: Erlbaum.

Costa, P. T., & McCrae, R. R. (1992). *Revised NEO personality inventory (NEO-PI-R) and NEO Five-Factor inventory (NEO-FFI)*. Odessa, FL: Psychological Assessment Resources.

Costa, P. T., Terracciano, A., & McCrae, R. R. (2001). Gender differences in personality traits across cultures: Robust and surprising findings. *Journal of Personality and Social Psychology*, *81*(2), 322–331. doi:10.1037//0022-3514.81.2.322

Feingold, A. (1994). Gender differences in personality: A meta-analysis. *Psychological Bulletin*, *116*(3), 429–456. doi:10.1037//0033-2909.116.3.429

Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum.

Le, L. T. (2009). Investigating gender differential item functioning across countries and test languages for PISA science items. *International Journal of Testing*, *9*(2), 122–133. doi:10.1080/15305050902880769

Marsh, H. W. (1996). Positive and negative global self-esteem: A substantively meaningful distinction or artifactors? *Journal of Personality and Social Psychology*, *70*(4), 810–819. doi:10.1037/0022-3514.70.4.810

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174. doi:10.1007/BF02296272

Mitchelson, J. K., Wicher, E. W., LeBreton, J. M., & Craig, S. B. (2009). Gender and ethnicity differences on the Abridged Big Five Circumplex (AB5C) of personality traits: A differential item functioning analysis. *Educational and Psychological Measurement*, *69*(4), 613–635. doi:10.1177/0013164408323235

Ostendorf, F., & Angleitner, A. (2004). *NEO-PI-R: NEO-Persönlichkeitsinventar nach Costa und McCrae*. Göttingen, Germany: Hogrefe.

Reise, S. P., Smith, L., & Furr, R. M. (2001). Invariance on the NEO-PI-R neuroticism scale. *Multivariate Behavioral Research*, *36*(1), 83–110. doi:10.1207/S15327906MBR3601_04

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, *14*(3), 271–282. doi:10.1177/014662169001400305

Rost, J. (1991). A logistic mixture distribution model for polytomous item responses. *The British Journal for Mathematical and Statistical Psychology*, *44*, 75–92. doi: 10.1111/j.2044-8317.1991.tb00951.x

Rost, J., Carstensen, C. H., & Davier, M. von. (1999). Sind die Big Five Rasch-skalierbar? - Eine Reananlyse der NEO-FFI-Normierungsdaten [Are the Big Five Rasch scalable? A re-analysis of the NEO-FFI norming data]. *Diagnostica*, *45*(3), 119–127. doi:10.1026//0012-1924.45.3.119

Rost, J., & Davier, M. von. (1995). Mixture distribution Rasch Models. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch Models: Foundations, recent developments, and applications* (pp. 257–268). New York, NY: Springer.

Smith, L., & Reise, S. P. (1998). Gender differences on negative affectivity: An IRT study of differential item functioning on the Multidimensional Personality Questionnaire stress reaction scale. *Journal of Personality and Social Psychology*, *75*(5), 1350–1362.

von Davier, M. (2001). WINMIRA 2001 [Computer software]. Kiel: Institute for Science Education.

Wang, W.-C. (2008). Assessment of differential item functioning. *Journal of Applied Measurement, 9(4),* 387-408.

Wetzel, E., Carstensen, C. H., & Böhnke, J. R. (in press). Consistency of extreme response style and non-extreme response style across traits. *Journal of Research in Personality*. Advance online publication. dx.doi.org/10.1016/j.jrp.2012.10.010

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. (2007). ConQuest (Version 2.0) [Computer software]. Camberwell, Australia: Australian Council for Educational Research.

Zieky, M. (1993). Practical questions in the use of DIF statistics in item development. In P. W. Holland & H. Wainer (Ed.), *Differential item functioning* (pp. 337–364). Hillsdale, NJ: Lawrence Erlbaum.

2.2. Appendix B: Manuscript Wetzel, Carstensen, & Böhnke (2013)

Journal of Research in Personality 47 (2013) 178–189

Contents lists available at SciVerse ScienceDirect

Journal of Research in Personality

journal homepage: www.elsevier.com/locate/jrp

Consistency of extreme response style and non-extreme response style across traits

Eunike Wetzel [a,*], Claus H. Carstensen [a], Jan R. Böhnke [b]

[a] Department of Psychology and Methods of Educational Research, Otto-Friedrich-University Bamberg, D-96045 Bamberg, Germany
[b] Clinical Psychology and Psychotherapy, University of Trier, D-54296 Trier, Germany

ARTICLE INFO

Article history:
Available online 2 November 2012

Keywords:
Response styles
Extreme response style
Non-extreme response style
Mixed Rasch models
Second order latent class analysis

ABSTRACT

The consistency of extreme response style (ERS) and non-extreme response style (NERS) across the latent variables assessed in an instrument is investigated. Analyses were conducted on several PISA 2006 attitude scales and the German NEO-PI-R. First, a mixed partial credit model (PCM) and a constrained mixed PCM were compared regarding model fit. If the constrained mixed PCM fit better, latent classes differed only in their response styles but not in the latent variable. For scales where this was the case, participants' membership to NERS or ERS on each scale was entered into a latent class analysis (LCA). For both instruments, this second order LCA revealed that the response style was consistent for the majority of the participants across latent variables.

1. Introduction

The aim of this paper is to analyze the consistency of response styles across the different latent variables assessed in an instrument. Response styles occur in many questionnaires employing Likert-type scales. However, it is unclear whether participants use the same response style throughout the instrument, independently of the trait being assessed, or whether there is a relationship between the occurrence of response styles and the trait. This paper tries to elucidate the consistency of response styles using mixed Rasch models, in which participants are allocated to response style classes for each of the scales, and a latent class analysis, in which the consistency of membership to a certain response style class is investigated. In the following, first a definition of response styles is provided and the importance of considering response styles is addressed. Second, the use of mixed Rasch models to investigate response styles is explained. Third, existing research on the stability and consistency of response styles is summarized and our approach to investigating the consistency of response styles is described.

1.1. Response styles in questionnaires

The term *response style* refers to systematic individual differences in response scale use that are independent of item content and the respondent's trait level. Thus, an individual's response style characterizes his or her tendency to prefer certain response categories over others. Response styles that have been shown to occur frequently are acquiescence response style, the tendency to agree with items, disacquiescence response style, the tendency to disagree with items, extreme response style (ERS), the tendency to prefer extreme response categories, and midpoint responding, the tendency to choose the middle category of a response scale (see Baumgartner and Steenkamp (2001) for a detailed summary of common response styles). Importantly, in all cases, the response style tendency is characterized by its occurrence irrespective of the item's content and the person's standing on the trait being assessed by the item.

The pervasiveness of these response styles has been shown in a wide variety of self-report questionnaires using Likert-type response scales. For example, Rost, Carstensen, and von Davier (1997) and Austin, Deary, and Egan (2006) found ERS in the German and English NEO-FFI (Borkenau & Ostendorf, 1993; Costa & McCrae, 1992) using mixed Rasch models. Eid and Rauber's (2000) mixed Rasch analysis of a leadership performance scale resulted in two subgroups of participants, one that preferred extreme categories and one that used the response scale evenly. Buckley (2009) showed the occurrence of acquiescence response style, disacquiescence response style, extreme response style, and non-contingent responding (i.e., inconsistent responses to similar items) in several attitude scales included in the Programme for International Student Assessment (PISA) 2006 student questionnaire (OECD, 2006).

Concerning the relationship between response styles and personality, Austin et al. (2006) found that persons employing ERS had higher extraversion and conscientiousness scores as measured by the NEO-FFI (however, see Paulhus, 1991). Naemi, Beal, and Payne (2009) showed that ERS and peer-ratings of intolerance of ambiguity, simplistic thinking, and decisiveness were positively

* Corresponding author. Address: Department of Psychology and Methods of Educational Research, Otto-Friedrich-University Bamberg, D-96045 Bamberg, Germany.
E-mail address: eunike.wetzel@uni-bamberg.de (E. Wetzel).

0092-6566/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jrp.2012.10.010

associated. Individual differences in response styles appear to be influenced by other factors as well. For example, Eid and Rauber (2000) reported that women had a higher probability of being allocated to the ERS group compared to men. Van Herk, Poortinga, and Verhallen (2004) showed that the occurrence of ERS and acquiescence response style differed between six European countries. Both response styles were more pronounced in Mediterranean than in Northwestern Europe. Johnson, Kulesa, Cho, and Shavitt (2005) analyzed data from 19 countries and found that the cultural dimensions power distance and masculinity were positively related to ERS whereas they were negatively related to acquiescence response style. Smith (2004) showed a relationship between acquiescence response style and nations that are high on family collectivism.

This paper focuses on response styles that differ in the degree of extremity of the preferred response. These are ERS and its opposite, namely a response style characterized by the avoidance of extreme response categories, called non-extreme response style (NERS) in the following, as well as midpoint responding. Note that NERS is not the same as midpoint responding, since midpoint responding is defined as an explicit preference for the middle category, while respondents employing NERS can prefer all moderate categories, including but not limited to the middle category. With the widely used response format *strongly disagree – disagree – neutral – agree – strongly agree* we would therefore expect ERS respondents to favor *strongly disagree* and *strongly agree*, midpoint responders to favor *neutral*, and NERS responders to favor either one of the moderate categories *disagree*, *neutral*, or *agree* irrespective of their true trait level.

The importance of considering response styles is illustrated by Austin et al. (2006) who showed that sum scores may be distorted when participants employ different response styles. In particular, ERS participants received more extreme trait scores compared to other participants. Thus, comparisons of sum scores across subgroups of participants may be rendered invalid by response styles. Furthermore, as Buckley (2009) pointed out, when attitudinal data obtained from international educational assessments such as PISA are used in secondary analyses, conclusions may be erroneous if individual and cross-cultural differences in response styles are not taken into account. Thus, the aim of this paper is to explore the consistency of response styles across traits in an instrument using several attitude scales from the PISA 2006 student questionnaire and a widespread personality questionnaire, the NEO-PI-R.

1.2. Identification of response styles using mixed Rasch models

In line with other studies on response styles (e.g., Austin et al., 2006; Eid & Rauber, 2000; Rost et al., 1997), mixed Rasch models (Rost, 1990; Rost, 1991) will be used to identify subgroups of participants that differ regarding their response style. Mixed Rasch models combine latent class analysis (LCA) with Rasch models. Both qualitative differences between subgroups of participants (as in an LCA) as well as quantitative differences between participants within a subgroup (as in a Rasch model) can be analyzed simultaneously. Thus, in a mixed Rasch model the Rasch model holds within each latent class but item parameters vary between latent classes.

The least restrictive mixed model for polytomous data (mixed partial credit model; Rost, 1991) extends the partial credit model (PCM; Masters, 1982) by incorporating class-specific parameters. According to Rost (1990, 1991, notation from Rost, 2004), it is defined as

$$P(X_{vi}=x)=\sum_{g=1}^{G}\pi_g\frac{\exp(x\theta_{vg}-\sigma_{ixg})}{\sum_{s=0}^{m}\exp(s\theta_{vg}-\sigma_{isg})} \quad (1)$$

with $\sigma_{ixg}=\sum_{s=1}^{x}\tau_{isg}$ for all ixg and two side conditions, $\sum_{i=1}^{k}\sum_{x=1}^{m}\tau_{ixg=0}$ for all g, and $\sigma_{i0g}=0$ for all i and g (explanation below). The probability of person v endorsing category x on item i with $m+1$ categories ($x=0,\ldots,m$) is denoted by $p(X_{vi}=x)$. The class size is given by π_g ($0<\pi_g<1$) with the constraint $\sum_{g=1}^{G}\pi_g=1$ for all g with the latent classes (g) being mutually exclusive.

In Eq. (1), θ_{vg} is the individual person parameter indicating the trait level of person v in latent class g. In analogy to the item difficulty in the dichotomous Rasch model, the item location in the PCM, which can be computed as the mean of the thresholds (see below), can be interpreted as the mean endorsement difficulty of the item. Thus, items with higher item locations are more difficult to endorse (i.e., higher trait levels are necessary to endorse response categories stating agreement) and items with lower item locations are easier to endorse. To illustrate, the gray line in Fig. 1a shows the item locations for the five items on the PISA 2006 student questionnaire scale *instrumental motivation in science*. As in the dichotomous Rasch model (Rasch, 1960), item location and person parameters are represented on the same latent trait with scale values in units of logits (depicted on the y-axis), which usually range between −3 and 3 (Embretson & Reise, 2000). For instance, in Fig. 1a, item 1 has a lower item location parameter (i.e., higher endorsement probability; 0.38 logits) than item 2 (0.96 logits).

Threshold parameters govern the responses in each item's categories. The threshold parameters indicate at which trait level it is equally likely for a respondent to answer in two adjacent categories. For implementation into the model in Eq. (1), σ_{ixg}, these threshold parameters are cumulated into item parameters $\sigma_{ixg}=\sum_{s=1}^{x}\tau_{isg}$ over all thresholds the participant's response x exceeded. In Fig. 1a, the black lines show the three threshold parameters for each of the five items on instrumental motivation in science. For example, the solid black line in Fig. 1a is the threshold between categories 1 (*strongly disagree*) and 2 (*disagree*). For item 1 it is located at about −4.2 logits. Since thresholds and trait values are estimated on the same logit scale, this indicates that a person with a trait value of −4.2 logits is equally likely to choose either *strongly disagree* or *disagree*. Likewise, respondents with trait values between two thresholds are most likely to respond in the corresponding category: respondents with trait values between −4.2 and about 0.5 (threshold 2 in Fig. 1a) will most likely choose *disagree*. Considering the probabilistic nature of the model these are always only statements about the most likely propensity for each person.

With the norming condition $\sum_{i=1}^{k}\sum_{x=1}^{m}\tau_{ixg}=0$ within each class g, effectively the mean of all item locations within each class is set to 0. A further condition ($\sigma_{i0g}=0$ for all i and g; Rost, 1991) allows the index for the response categories x to be used for the notation of the thresholds τ_{isg} as well.

In a mixed PCM with more than one latent class ($g>1$ in Eq. (1)), the PCM holds within each latent class but item parameters may be different between the classes. Item parameter invariance between samples is a property of unidimensional traits and, for example, is subject to testing the homogeneity of scales (Andersen, 1973). With item location and threshold parameters being different between classes, the latent variables measured in such classes strictly speaking have different meanings, i.e., different traits are measured in latent classes with differing item parameters. The differences between latent classes can be interpreted as content-related differences (e.g., different traits are being measured) as well as content-unrelated differences (e.g., differences in response scale usage).

For the examination of the consistency of response styles, it is desirable to ensure that participants solely differ in their response scale use on the scales under investigation, but not in the trait that is being assessed, their understanding of the items' content, or

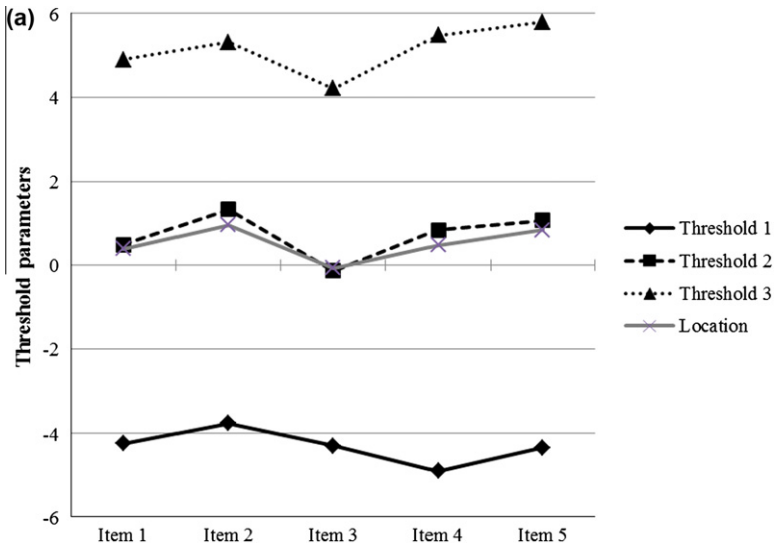

Fig. 1a. Threshold parameters for NERS on instrumental motivation in science under the constrained mixed partial credit model.

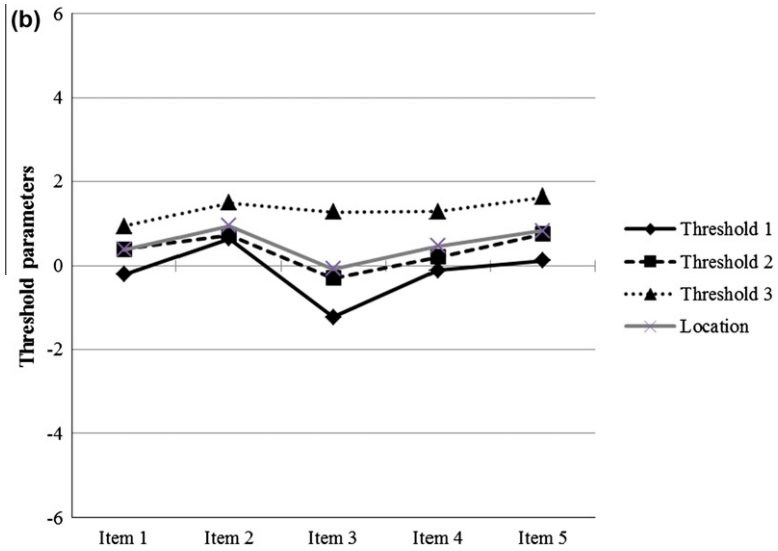

Fig. 1b. Threshold parameters for ERS on instrumental motivation in science under the constrained mixed partial credit model.

other factors that might influence the choice of a response category. The central idea of the approach presented in this paper is to differentiate between differences in item locations, which are interpreted as capturing different traits, and differences between classes in threshold parameters, which reflect different response styles while responses can be assumed to be on the same latent trait. Whether the latent classes are homogeneous regarding the trait being measured and only differ in response styles can be tested by model comparisons between a regular mixed PCM as described above and a constrained mixed PCM. Instead of estimating all parameters (locations, thresholds) freely for each class as in the unconstrained mixed PCM, for the constrained mixed PCM, item locations are restricted to be equal between latent classes yielding $\sigma_{ixg} = \sigma_{ix}$ in the model in Eq. (1).

Since all parameters are estimated freely in the unconstrained mixed PCM, the resulting latent classes can differ regarding response styles as well as other factors such as the trait being measured. With the equality constraint imposed on the item location parameters in the constrained mixed PCM, homogeneous latent classes can be assumed which can only differ in the distribution of the threshold parameters τ_{ixg} characterizing different response styles. This is illustrated in Fig. 1 which shows the characteristic difference in threshold parameters between NERS and ERS for the PISA 2006 attitude scale instrumental motivation in science. For the NERS group (Fig. 1a), thresholds are widely spaced while for the ERS group (Fig. 1b), the three thresholds are close together. Due to the equality constraint implemented in the constrained mixed PCM, the location parameters (grey lines in Fig. 1) are the same for both classes. Thus, the classes only differ in the distribution of their threshold parameters. For participants allocated to the NERS group, the trait level necessary to choose one of the outer categories (*strongly disagree* or *strongly agree*) is more extreme than for participants allocated to the ERS group. For example, on item 5 of instrumental motivation in science, a NERS person would need a trait value of about 6 for *strongly agree* to be the most likely category, while for an ERS person a trait value of about 2 would suffice. Thus, participants in the NERS group can be interpreted as respondents who prefer middle categories while participants in the ERS group can be interpreted as participants who prefer extreme categories.

If the constrained mixed PCM holds for observed data, confirming trait homogeneity between the latent classes, trait values are directly comparable between latent classes. By constraining item location parameters to be equal it is ensured that trait values are on the same scale while potential differences in response style use are captured by the threshold parameters. Thus, trait values based on the constrained mixed PCM are corrected for response styles (see Rost et al., 1997). Since sum scores may be affected by different response styles, only trait values from a constrained mixed PCM should be used to compare the trait levels of respondents from different latent classes (i.e., response styles).

In sum, the approach taken in this paper to operationalize response styles is to compare a mixed PCM and a constrained mixed PCM regarding model fit for the scales under investigation. If the assumption of trait homogeneity between the latent classes holds, they only differ with respect to their response scale usage. Scales in which this is the case will be included in the analysis of the consistency of response styles across traits presented below.

1.3. Stability and consistency of response styles

Several studies have explored the stability of response styles longitudinally and across traits. Regarding the longitudinal stability of response styles, Bachman and O'Malley (1984) reported high reliability estimates for an agreement and an extreme responding index for a follow-up period of up to four years across five questionnaire forms. Participants in Weijters, Geuens, and Schillewaert's (2010b) study filled out two different online questionnaires with a one-year interval between data collections. Weijters, Geuens et al. (2010b) analyzed the stability of four response styles (acquiescence response style, disacquiescence response style, extreme response style, and midpoint responding) using a second order measurement model that included time-specific response style factors for the two waves and second order time-invariant response style factors. They found that more than half of the variance in the time-specific response style factors was explained by their respective time-invariant response style factor, supporting a high stability of the four response styles over a one-year period.

Concerning the consistency of response styles across traits within a questionnaire, Austin et al. (2006) found that membership to either the ERS or NERS latent class in (unconstrained) mixed Rasch models correlated significantly and positively between neuroticism, extraversion, agreeableness, and conscientiousness, indicating that participants applied the same response style over the course of the NEO-FFI. Similarly, using correlations between class memberships derived from mixed Rasch models as well, Hernández, Drasgow, and González-Romá (2004) reported that about 49% of their participants were consistently allocated to the class avoiding the middle category across the traits assessed in the 16PF Questionnaire (Cattell, Cattell, & Cattell, 1993), though participants demonstrating a preference for the middle category did not do so consistently. Furthermore, Weijters, Geuens, and Schillewaert (2010a) used structural equation modeling to show that acquiescence response style and extreme response style were mostly consistent across a random sample of items taken from marketing and attitude scales. They found that response styles were best modeled using a tau-equivalent factor model with a time-invariant autoregressive coefficient, indicating that the effect of the two response styles generalized across independent item sets.

In this paper, we take an alternative approach to testing the consistency of response styles across the traits assessed in a questionnaire, namely a second order latent class analysis (Keller & Kempf, 1997). That is, mixed Rasch models will be computed first to allocate participants to different response styles. Then, a latent class analysis will be computed using the response style assignments resulting from the mixed Rasch models. In the following, analyses conducted on several PISA 2006 attitude scales (study 1) and the NEO-PI-R (study 2) will be reported. The results from both studies will be discussed in the general discussion.

2. Study 1: consistency of ERS and NERS in the PISA 2006 attitude scales

2.1. Method

2.1.1. Sample

In study 1, data from the German students taking part in the PISA 2006 assessment (''PISA sample'') was analyzed.[1] Only the German PISA 2006 sample (as opposed to all the countries taking part in PISA 2006) was used to avoid cross-cultural differences in response styles (e.g., Johnson et al., 2005) from contaminating the analyses. For an investigation of cultural differences in response styles using PISA 2006 data from 57 countries see Buckley (2009). In total, 4891 (49.1% girls) between the ages of 15 and 16 (M = 15.85, SD = 0.28) formed the PISA sample. To be able to validate the results, the sample was randomly split into halves.

2.1.2. Instrument

PISA is a triennial educational large scale assessment of 15-year-olds in the domains reading, mathematics, and science conducted by the Organisation for Economic Cooperation and Development (OECD). While the focus is on the cognitive tests, students additionally fill out a student questionnaire. The student questionnaire contains questions regarding the students' backgrounds (e.g., their parents' occupations) as well as several scales assessing the students' attitudes towards the PISA competency domains, towards learning, and towards their future-related motivation to enter a career in one of the domains (OECD, 2006). In PISA 2006, with science being the major domain, several science-specific scales concerning students' learning strategies, motivations, and self-concept in science were included. In this study, nine attitude scales were used for the analyses, the first eight in Table 1 being part of the international student questionnaire and the ninth, *reading enjoyment*, taken from the national (German) student questionnaire additionally filled out by the German PISA sample (Frey et al., 2009). These scales were chosen because they share the same response format (*strongly disagree – disagree – agree – strongly agree*), which is also comparable to the one used in the NEO-PI-R (see study 2), while the other attitude scales have different response scales. Table 1 shows the number of items in each scale, means, and standard deviations of the weighted likelihood estimates (WLE; Warm, 1989) for the sample on the nine scales as well as Cronbach's α reliabilities for the test scores.

2.1.3. Analyses

2.1.3.1. Pre-analyses: identification of response styles. To identify subgroups of participants who differed regarding their response styles, the data were analyzed using the mixed PCM (Rost, 1991). For each of the PISA 2006 attitude scales, the mixed PCM and the constrained mixed PCM with two latent classes each were compared regarding model fit. WINMIRA (von Davier, 2001) applies conditional maximum likelihood estimation using an EM algorithm (Bock & Aitkin, 1981) to estimate the item parameters. The score distributions are by default approximated using a logistic distribution (*smoothed score distribution*). For scales in which the approximation was not sufficiently close to the observed distribution (RMSEA > .08), the unconstrained score distribution was estimated. The better-fitting model was determined using the

[1] Public use data from the PISA assessments is available online at http://www.oecd.org/pages/0,3417,en_32252351_32235731_1_1_1_1_1,00.html.

Table 1
Descriptive statistics for the PISA 2006 attitude scales.

Scale	Number of items	WLE mean (SD)	Cronbach's α
Science enjoyment	5	−.08 (1.09)	.92
General science value	5	−.09 (1.06)	.75
Personal science value	5	−.23 (1.07)	.81
Environment responsibility	7	.08 (.94)	.76
Usefulness for science career	4	.11 (1.09)	.83
Future-oriented science motivation	4	−.15 (1.02)	.91
Instrumental motivation in science	5	−.08 (1.04)	.90
Science self-concept	6	.26 (.99)	.90
Reading enjoyment	9	.04 (.74)	.91

Note: WLE = weighted likelihood estimate.

Consistent Akaike's Information Criterion (CAIC; Bozdogan, 1987), an information criterion that takes sample size into account, penalizes overparameterization, and is asymptotically consistent. Only scales in which the constrained mixed PCM showed a better fit (indicated by a lower CAIC value) were included in the following analyses of the consistency of response styles across traits.

In the constrained mixed PCMs, each person was allocated to one of the latent classes based on their maximum probability of class membership. That is, separately for every scale, the probability of being member of each latent class was estimated for every person and the person was assigned member of the latent class with the highest probability. The mean of these probabilities across all scales (called mean maximum probability of class membership) was used to evaluate whether participants could be allocated to one class with high certainty (high mean maximum probability) or whether allocations were ambiguous (low mean maximum probability). The latent classes were then interpreted as response styles using the threshold parameters. To investigate the impact of response styles on trait values, correlations between the scales were compared using the trait values resulting from a one-class standard PCM (no distinction of response styles) and the trait values derived from constrained mixed PCMs which are estimated separately for each latent class and thus take response styles into account. Furthermore, correlations between standard PCM trait values and corrected constrained mixed PCM trait values within one scale were considered.

2.1.3.2. Consistency of response styles across traits. The consistency with which participants were allocated to a certain response style class across scales was assessed using a latent class analysis (LCA) on the (manifest) class membership data in WINMIRA. That is, the class memberships assigned to participants in the constrained mixed PCMs for each scale were entered as variables in an LCA. The number of classes appropriate for this second order LCA was determined by the CAIC (Bozdogan, 1987).

2.1.4. Validation

To validate the results obtained from the first half of the German PISA 2006 sample, analyses were repeated using the second half. The procedure was identical to the one described above for the first half.

2.2. Results

2.2.1. Pre-analyses: identification of response styles

Table 2 shows model fit comparisons between the mixed PCM and the constrained mixed PCM for the PISA 2006 attitude scales. For all of the attitude scales analyzed here the two-class mixed PCMs fit better compared to the one-class PCMs. Concerning the two-class models for two scales, namely *future-oriented science motivation* and *reading enjoyment*, the mixed PCM yielded a lower CAIC value than the constrained mixed PCM, indicating trait heterogeneity between the classes. Thus, the second order LCA was computed without these two scales. The latent classes were interpreted as NERS and ERS using the threshold parameters (see Fig. 1 and explanation in the introduction for an illustration). The characteristic response patterns of NERS and ERS are also apparent in the different category frequencies. To illustrate, Fig. 2 shows the relative frequency of each response category for each of the five *science enjoyment* items, separately for members of the NERS and ERS classes. As can be seen in this figure, for NERS, *disagree* and *agree* were the most frequently chosen response options whereas for ERS, *strongly disagree* and *strongly agree* were the most frequently chosen response options. The mean maximum probability of class membership assignment was .87 (*SD* = .07) for the seven attitude scales. Thus, participants could be allocated to one of the response style classes with high certainty.

To examine whether trait values were influenced by response styles, trait value correlations between the attitude scales for a standard PCM and the constrained mixed PCM (corrected for response styles) were compared. As Table 3 shows, correlations were

Table 2
Comparison of model fit for the mixed partial credit model (mixed PCM) and the constrained mixed PCM for the PISA 2006 attitude scales.

Scale	Mixed PCM LL	*N*par	Mixed PCM CAIC	Constrained mixed PCM LL	*N*par	Constrained mixed PCM CAIC	*N*
Science enjoyment	−10667.71	33	21624.12	−10669.27	28	**21583.50**	2318
General science value[a]	−11102.65	57	22702.34	−11112.12	52	**22677.67**	2253
Personal science value	−12499.53	33	25286.69	−12507.42	28	**25258.90**	2245
Environment responsibility[a]	−16010.03	81	32725.34	−16028.71	74	**32701.75**	2224
Usefulness for science career	−8784.14	27	17803.55	−8787.85	23	**17776.12**	2239
Future-oriented science motivation[a]	−8097.76	45	**16587.74**	−8125.00	41	16607.36	2244
Instrumental motivation in science	−11085.04	33	22457.41	−11087.78	28	**22419.34**	2223
Science self-concept	−11881.45	39	24102.27	−11894.23	33	**24075.61**	2212
Reading enjoyment	−19355.35	57	**39204.71**	−19540.94	48	39497.89	2136

Note: LL = log-likelihood, Npar = number of parameters, CAIC = Consistent Akaike's Information Criterion. The CAIC of the better-fitting model is depicted in boldface.
[a] Facets for which the score distribution in WINMIRA was not approximated.

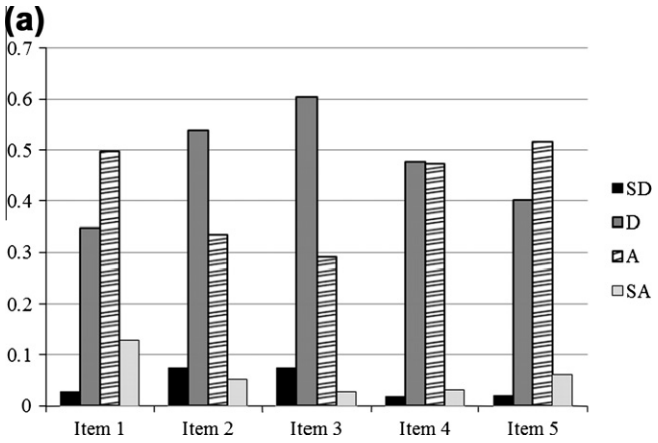

Fig. 2a. Category frequencies for NERS (class size 63.3%) on science enjoyment

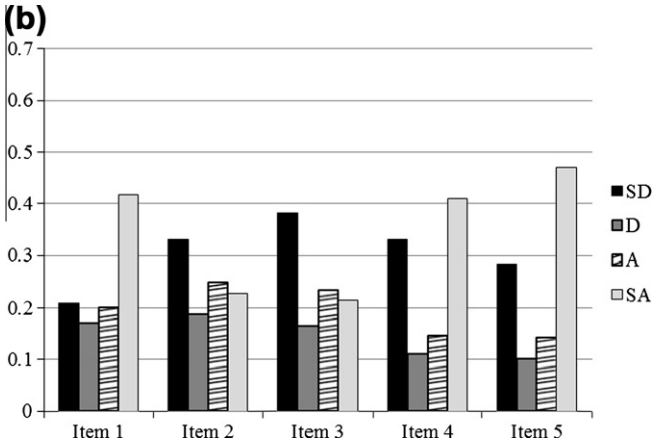

Fig. 2b. Category frequencies for ERS (class size 36.7%) on science enjoyment. SD = strongly disagree, D = disagree, A = agree, SA = strongly agree.

consistently higher for the standard PCM compared to the constrained mixed PCM correlations. Standard PCM trait values and corrected constrained mixed PCM trait values correlated highly, but not perfectly with each other (see diagonal in Table 3).

2.2.2. Consistency of response styles across traits

The second order LCA was computed using the class membership assignments on the seven scales with better-fitting constrained mixed PCMs. According to the CAIC, a two-class solution yielded the best fit to the data (Table 4). A 69.1% of the participants were allocated to the first latent class and 30.9% to the second. The mean probability for class membership allocation was .90 (*SD* = .13), indicating a high certainty in the assignment of participants to the two latent classes. Fig. 3 shows the category probabilities for being assigned to NERS or ERS on the seven PISA 2006 attitude scales. For class 1, probabilities are highest for assignment to the NERS group across all seven scales. Thus, participants allocated to class 1 in the second order LCA appear to use NERS consistently across the seven latent variables being assessed. For class 2, a different picture emerges. Here, the ERS category is the most probable one for all scales with the exception of *usefulness for science career*, for which NERS still has the highest probability. Note however, that category probabilities on average are lower compared to class 1. Hence, class 2 appears to contain participants who by tendency used ERS consistently as well as participants who switched between the two response styles.

2.2.3. Validation

As in the first half of the PISA sample, the two-class models in general showed a superior fit compared to the one-class models. The model comparison between mixed PCM and constrained mixed PCM for the second half of the PISA sample yielded eight attitude scales for which the constrained mixed PCM showed a better fit. In addition to the scales remaining for the first half of the PISA sample, the constrained mixed PCM also fit better for future-oriented science motivation. The two latent classes were identified as NERS and ERS using the threshold parameters. The mean maximum probability for class membership assignment was .87 (*SD* = .06). The second order LCA using the class membership on the eight remaining attitude scales resulted in a two-class solution (see Table 4). The two latent classes could be characterized similarly to the first half of the PISA sample concerning the common seven scales. Class 1 (64.1%) contained consistent non-extreme responders. In class 2 (35.9%), ERS was the most probable category for the majority of the scales, though for usefulness for science career and future-oriented science motivation, NERS had the highest category probability.

3. Study 2: consistency of ERS and NERS in the NEO-PI-R

3.1. Method

3.1.1. Sample

The sample in study 2 (“NEO sample”) consisted of the non-clinical standardization sample (*N* = 11,724; 64.0% women) for the German NEO-PI-R (Ostendorf & Angleitner, 2004). Participants were between 16 and 91 years old (*M* = 29.92, *SD* = 12.08). The sample was randomly divided into two halves, allowing the results obtained using the first half to be validated with the second half.

Table 3
Trait value correlations between PISA attitude scales for the partial credit model (above diagonal) and the constrained mixed partial credit model (below diagonal).

Scale	SCJOY	GENVAL	PERVAL	ENVRES	USECAR	INSCMO	SCSELF
Trait value correlations							
SCJOY	**.923**	.482	.639	.323	.288	.479	.508
GENVAL	.383	**.801**	.631	.290	.268	.338	.332
PERVAL	.569	.509	**.924**	.304	.299	.533	.439
ENVRES	.296	.230	.279	**.945**	.184	.197	.280
USECAR	.257	.215	.245	.135	**.826**	.363	.323
INSCMO	.439	.266	.477	.184	.299	**.890**	.471
SCSELF	.478	.249	.401	.240	.246	.423	**.898**

Note: **Diagonal** (bold): correlations between standard partial credit model trait values and corrected constrained mixed partial credit model trait values within each scale. **Above diagonal**: correlations between PISA attitude scales based on standard trait values from the partial credit model. **Below diagonal**: correlations between PISA attitude scales based on corrected trait values from the constrained mixed partial credit model.
All correlations are significant at the $p = .001$ level. SCJOY = science enjoyment, GENVAL = general science value, PERVAL = personal science value, ENVRES = environment responsibility, USECAR = usefulness for science career, INSCMO = instrumental motivation in science, SCSELF = science self-concept.

Table 4
Results from the second order latent class analysis for the PISA 2006 attitude scales and the NEO-PI-R.

Scale and Sample	Classes	LL	Npar	CAIC	N
PISA 2006	1	−8819.82	7	17699.94	2025
Half 1	**2**	**−8375.39**	**15**	**16879.99**	**2025**
	3	−8341.81	23	16881.72	2025
	4	−8320.46	31	16907.94	2025
	5	−8306.57	39	16949.07	2025
	6	−8299.97	47	17004.77	2025
PISA 2006	1	−9612.87	8	19294.48	1986
Half 2	**2**	**−9091.87**	**17**	**18329.83**	**1986**
	3	−9060.34	26	18344.12	1986
	4	−9040.11	35	18381.01	1986
	5	−9028.61	44	18435.36	1986
	6	−9022.43	53	18500.34	1986
NEO–PI-R	1	−52096.31	18	104364.32	5109
Half 1	2	−45935.23	37	92223.39	5109
	3	−45009.56	56	90553.29	5109
	4	−44842.79	**75**	90400.99	**5109**
	5	−44764.25	94	90425.14	5109
	6	−44716.23	113	90510.34	5109
NEO-PI-R	1	−45323.41	15	90790.06	5165
Half 2	2	−40154.32	31	80604.68	5165
	3	−39482.73	47	79414.29	5165
	4	**−39351.73**	**63**	79305.09	**5165**
	5	−39277.21	79	79308.84	5165
	6	−39233.77	95	79374.76	5165

Note: LL = log-likelihood, Npar = number of parameters, CAIC = Consistent Akaike's Information Criterion The best-fitting class solution is in bold.

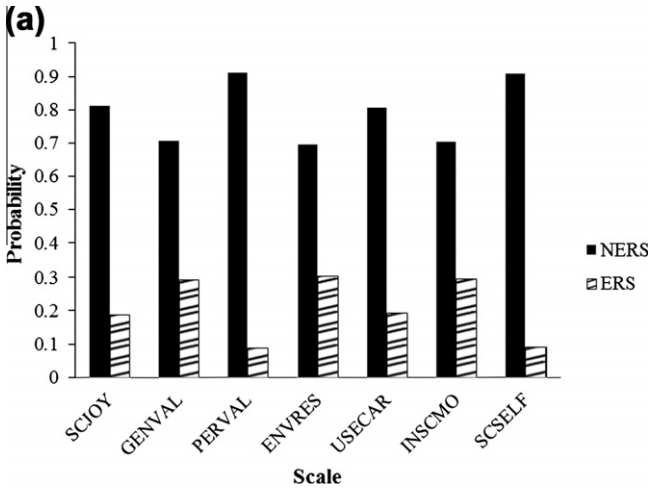

Fig. 3a. Category probabilities for the PISA attitude scales in Class 1 with size 69.1%.

3.1.2. Instrument

Participants filled out the German NEO-PI-R (Ostendorf & Angleitner, 2004). The NEO-PI-R assesses the Big Five personality domains neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness. Each of these higher-order domains consists of six facets, which are assessed by eight items each, adding up to 240 items in total. Responses are given on a five-point Likert-type scale with the response options *strongly disagree, disagree, neutral, agree,* and *strongly agree*. Cronbach's α reliabilities for sum scores on the facets ranged from .64 to .85 for neuroticism, .60–.80 for extraversion, and .53–.81 for openness to experience. For agreeableness, the range was .60–.76 and for conscientiousness it was .65–.81.

3.1.3. Analyses and validation

The analyses conducted in study 2 followed the same procedure as described above for study 1. First, for each facet, a mixed PCM

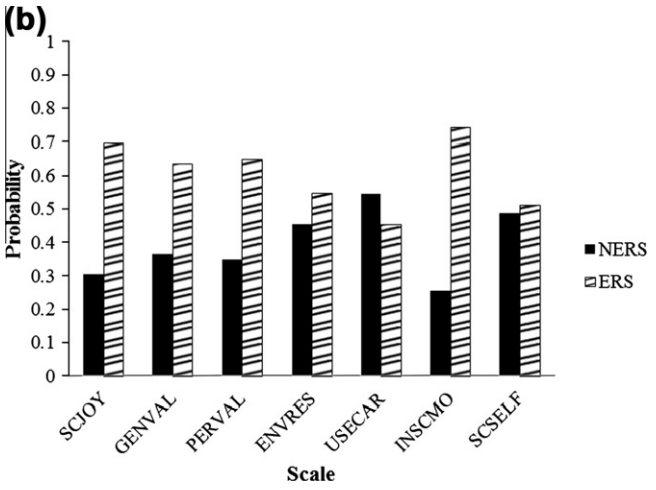

Fig. 3b. Category probabilities for the PISA attitude scales in Class 2 with size 30.9%. SCJOY = science enjoyment, GENVAL = general science value, PERVAL = personal science value, ENVRES = environment responsibility, USECAR = usefulness for science career, INSCMO = instrumental motivation in science, SCSELF = science self-concept.

and a constrained mixed PCM were compared concerning model fit. Drawing upon the results from other studies analyzing a NEO questionnaire with mixed Rasch models (Austin et al., 2006; Rost et al., 1997; Wetzel, Böhnke, Carstensen, Ziegler, & Ostendorf, 2012) two-class and three-class solutions were estimated. Response styles were subsequently identified using the threshold parameters. Furthermore, correlations of trait values on the NEO facets were compared between a one-class PCM that does not take response styles into account and the constrained mixed PCMs which differentiate between response styles. Second, a second order latent class analysis was computed using the class membership assignments from the constrained mixed PCMs. To validate the results obtained from the first half of the standardization sample, analyses were repeated using the second half. The procedure was identical to the one described above.

3.2. Results

3.2.1. Pre-analyses: identification of response styles

To identify latent classes that only differed regarding their response scale usage, the fit of a mixed PCM and a constrained mixed PCM were compared for each facet (see Table 5). Five facets had to be removed due to the occurrence of null categories which cause estimation problems. Over the remaining 25 NEO-PI-R facets, the constrained mixed PCM resulted in a lower CAIC value for 16 facets. Fourteen of these were two-class solutions while *openness to actions* and *deliberation* yielded three-class solutions. The latent classes in these 16 facets were interpreted as response styles using the threshold parameters as described in the introduction. The third class emerging for openness to actions and deliberation was very similar to NERS in that the first and fourth thresholds were widely spaced. It differed from NERS in the use of the middle category: While the first NERS class used the middle category *neutral* rarely (nearly overlapping second and third threshold), the second NERS class (NERS 2) did not appear to use the middle category at all (reversed and widely spaced second and third thresholds). The assignment of participants to either NERS, NERS 2, or ERS had a mean maximum probability of class membership of .86 (*SD* = .05). Exemplarily for neuroticism and extraversion, correlations between trait values resulting from a standard PCM and the constrained mixed PCM are contrasted in Table 6. In almost all cases, standard PCM correlations were higher than constrained

Table 5
Comparison of model fit for the mixed partial credit model (mixed PCM) and the constrained mixed PCM for the NEO-PI-R.

Comparison of Model Fit for the Mixed Partial Credit Model (mixed PCM) and the Constrained Mixed PCM for the NEO-PI-R **Facet**		**Mixed PCM LL**	**Mixed PCM CAIC**	**Constrained mixed PCM LL**	**Constrained mixed PCM CAIC**	**N**
Neuroticism						
N1 Anxiety		−57914.04	116475.55	−57937.38	**116444.92**	5789
N2 Angry hostility		−58524.91	117697.04	−58458.95	**117487.84**	5768
N3 Depression		−57060.57	**114768.79**	−57114.06	114798.43	5804
N4 Self-consciousness*		−60661.81	**122532.17**	−60709.60	122550.39	5816
N5 Impulsiveness		−60791.16	122229.78	−60829.54	**122229.25**	5789
N6 Vulnerability*		−52787.58	**106783.03**	−52854.33	106839.24	5785
Extraversion						
E1 Warmth		−51568.99	**103785.36**	−51664.50	103899.09	5782
E2 Gregariousness		−58637.78	117922.52	−58672.09	**117913.90**	5746
E3 Assertiveness		−57565.96	115779.34	−57603.21	**115776.54**	5785
E4 Activity		−58098.99	116845.55	−57873.30	**116316.85**	5797
E5 Excitement-seeking*		−66915.58	135039.28	−66943.99	**135018.78**	5796
E6 Positive emotions*		−54077.43	109362.84	nc	nc	5790
Openness to experience						
O1 Fantasy		−56058.18	**112763.46**	−56217.69	113005.22	5758
O2 Aesthetics*		−56524.36	**114256.96**	−56611.67	114354.27	5802
O3 Feelings		−50433.73	101515.02	−50460.44	**101491.12**	5797
O4 Actions*	2 classes	−59671.26	120551.17	−59690.41	120512.11	5821
	3 classes	−59373.21	120554.56	−59393.49	**120440.43**	5821
O5 Ideas*		−58149.13	**117506.70**	−58340.12	117811.35	5811
O6 Values		−58064.11	116775.60	−58075.17	**116720.44**	5782
Agreeableness						
A1 Trust		nc	nc	−55921.67	112413.47	5786
A2 Straightforwardness		−59550.77	**119749.15**	−59904.27	120378.83	5801
A3 Altruism*		nc	nc	nc	nc	5790
A4 Compliance*		−58627.70	118463.69	−58635.43	**118401.82**	5804
A5 Modesty		nc	nc	−58435.78	117441.59	5777
A6 Tender-mindedness		−54184.46	109016.41	−54203.30	**108976.77**	5790
Conscientiousness						
C1 Competence		−52383.08	105413.52	nc	nc	5779
C2 Order*		−59231.31	119671.13	−59268.22	**119667.59**	5814
C3 Dutifulness		−53637.04	**107921.57**	−53712.21	107994.60	5791
C4 Achievement striving*		−59358.01	119924.31	−59381.33	**119893.61**	5804
C5 Self-discipline		−55376.84	111401.10	−55390.49	**111351.11**	5786
C6 Deliberation*	2 classes	−56722.95	114654.30	−56736.63	114604.31	5809
	3 classes	−56124.80	114057.35	−56186.08	**114025.25**	5809

Note. Number of parameters for mixed PCM = 67, number of parameters for constrained mixed PCM = 59, LL = log-likelihood, CAIC = Consistent Akaike's Information Criterion, nc = null categories. The CAIC of the better-fitting model is depicted in boldface.
* Facets for which the score distribution in WINMIRA was not approximated. Number of parameters for mixed PCM = 125, number of parameters for constrained mixed PCM = 117.

mixed PCM correlations. Moreover, standard PCM trait values correlated highly, but not perfectly, with corrected constrained mixed PCM trait values for these seven facets.

3.2.2. Consistency of response styles across traits

Facets for which the constrained mixed PCM showed a better fit than the mixed PCM were included in the second order latent class analysis. This LCA was computed using the class membership variables obtained in the first order constrained mixed PCMs which allocated participants to the ERS or NERS (or for openness to actions and deliberation NERS 2) response style group. The second order LCA resulted in a four-class solution according to the CAIC (Table 4). The mean probability of allocation to one of the four latent classes was .85 (*SD* = .14). The category probabilities for the four classes revealed that for three of the classes, one of the categories (NERS (2) or ERS) was the most probable for all facets (see Fig. 4a, 4b, and 4d). For the first and second class, the highest probability was always the NERS or NERS 2 category while for the fourth class the highest probability was always the ERS category. Thus, for three classes (containing about 80% of the participants) the response style occurred consistently irrespective of which trait was being assessed. For the third class (about 20%), a different picture emerges. Here, probabilities for both categories were between .40 and .60 for the 14 facets with two-class solutions. For openness to actions and deliberation, probabilities were around .40 for ERS

Table 6
Trait value correlations between neuroticism and extraversion facets for the partial credit model (above diagonal) and the constrained mixed partial credit model (below diagonal).

Facet	N1	N2	N5	E2	E3	E4	E5
Trait value correlations							
N1	**.930**	.556	.257	−.095	−.328	−.084	−.088
N2	.515	**.935**	.373	−.090	−.063	.124	.065
N5	.245	.338	**.945**	.196	.068	.169	.272
E2	−.093	−.099	.180	**.944**	.304	.258	.437
E3	−.298	−.055	.078	.263	**.937**	.462	.248
E4	−.082	.100	.154	.219	.439	**.926**	.202
E5	−.081	.063	.245	.405	.229	.197	**.942**

Note: **Diagonal** (bold): correlations between standard partial credit model trait values and corrected constrained mixed partial credit model trait values within each scale. **Above diagonal**: correlations between NEO-PI-R facets based on standard trait values from the partial credit model. **Below diagonal**: correlations between NEO-PI-R facets based on corrected trait values from the constrained mixed partial credit model.
All correlations are significant at the $p = .001$ level. N1 = anxiety, N2 = angry hostility, N5 = impulsiveness, E2 = gregariousness, E3 = assertiveness, E4 = activity, E5 = excitement-seeking.

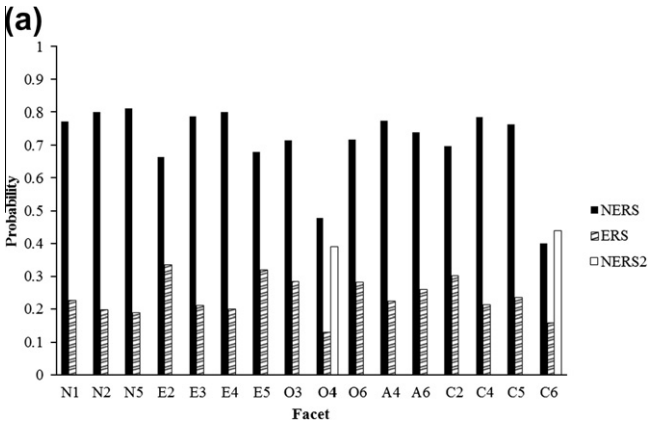

Fig. 4a. Category probabilities in NEO-PI-R Class 1 with size 39.2%.

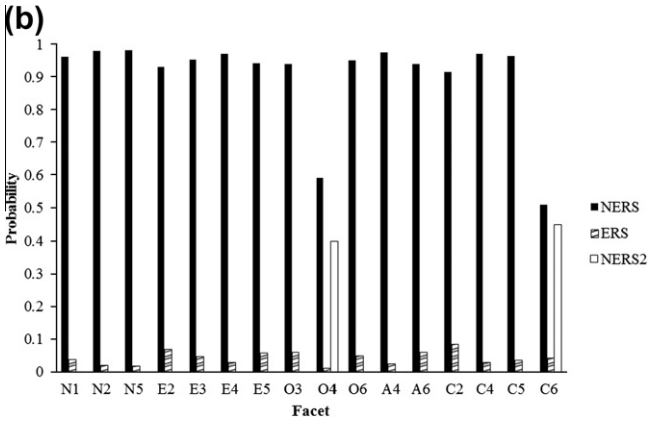

Fig. 4b. Category probabilities in NEO-PI-R Class 2 with size 35.8%.

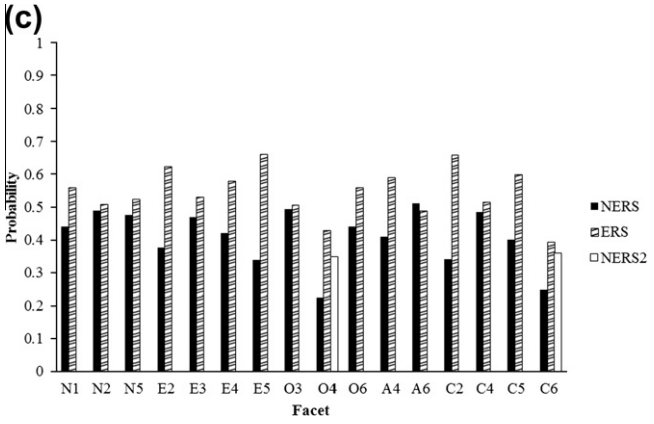

Fig. 4c. Category probabilities in NEO-PI-R Class 3 with size 19.6%.

and around .60 for a combined NERS and NERS 2 category (Fig. 4c). It follows that for participants allocated to this class, it was not possible to classify them clearly as belonging to one response style. Again there appears to be no contingency between traits and response styles.

3.2.3. Validation

For the second half of the NEO sample, the constrained mixed PCM showed a better fit compared to the mixed PCM on 14 facets. Eleven of these facets were the same between half 1 and half 2 of the NEO sample (N2, N5, E3, E4, E5, O3, O4, A4, C2, C5, and C6). Warmth (E1), trust (A1), and straight forwardness (A2) yielded better-fitting constrained mixed PCMs in the second half but not in the first half, while the reverse was the case for anxiety (N1), gregariousness (E2), openness to values (O6), tender-mindedness (A6), and achievement striving (4). For 12 of the 14 facets, two classes were appropriate to describe the data. These two classes were interpreted as NERS and ERS. Openness to actions again resulted in a three-class solution which could be interpreted as depicted for half 1 of the NEO sample, with the third class being non-extreme responders who did not employ the middle category (NERS 2). The mean maximum probability of class membership was .87 (*SD* = .05).

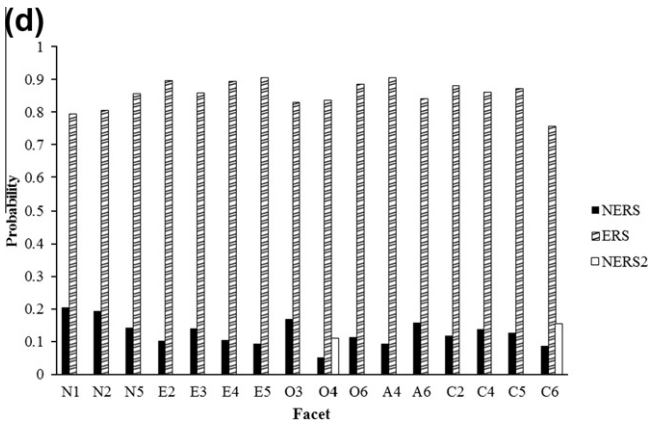

Fig. 4d. Category probabilities in NEO-PI-R Class 4 with size 5.4%.

The second order latent class analysis with the 14 remaining facets resulted in a four-class solution (Table 4) with a mean maximum probability of class membership of .82 (*SD* = .14). For the largest two classes (class 1 with 42.0% and class 2 with 28.1%) NERS was the most probable category across all facets. For class 3 (23.0%) probabilities for ERS and NERS (or combined NERS and NERS 2 for O4) were around 50% for ten facets, while for three (E1, E5, and C2) ERS was the most probable category. In class 4 (6.9%), ERS was the most probable category across all facets. In sum, for 77.0% (class 1, 2, and 4) of the participants the response style occurred consistently across traits. For 23.0% (class 3) the response style appears to be inconsistent.

4. General discussion

In this paper, the consistency of two response styles, NERS and ERS, was investigated across latent variables in several PISA 2006 attitude scales and the NEO-PI-R. For the majority of the participants in both instruments the response style occurred consistently independently of the trait that was being assessed. In the following, first the modeling of response styles suggested in this paper will be discussed. Then the implications of the occurrence and consistency of response styles in the two instruments and factors influencing the consistency of response styles will be addressed.

4.1. Modeling response styles

According to Baumgartner and Steenkamp (2001, p. 144), "the major problem in measuring response styles is not to confound stylistic variance with substantive variance". This problem is mostly dealt with by assessing response styles using an item set that is heterogeneous in content and computing an index for each response style. For example, Weijters, Cabooter, and Schillewaert (2010) compute ERS by taking the natural logarithm of the number of extreme responses plus one divided by the number of non-extreme

responses plus one. Weijters, Geuens et al. (2010a) recommend that the items used to compute the index should be a random sample of items from inventories assessing heterogeneous traits.

We suggest a different approach to operationalizing response styles, which does not require a trait-heterogeneous item set and instead is model-based. In our study, response styles were operationalized using a model comparison between a mixed partial credit model and a constrained mixed partial credit model. If two classes result in a mixed Rasch model, they do not necessarily differ regarding the construct being measured or the class members' understanding of the items. Instead, heterogeneity can also be due to differences in response style. For the subsequent analyses of response styles it is important to ensure that the latent classes derived from the mixed Rasch models differ only regarding their response style and style is therefore not confounded with substance. This can be achieved with the equality constraint implemented in the constrained mixed PCM. If the constrained mixed PCM, in which item location parameters are constrained to equality between the latent classes, shows a better fit than the mixed PCM, it can be concluded that the latent classes do not differ regarding the construct being assessed but instead differ in their response scale use, since that is the only difference the constrained mixed PCM allows. In contrast, if the mixed PCM holds for the data, the latent classes may differ regarding the trait that is being assessed. After the model comparison, the latent classes can be identified as different response styles using the separately estimated threshold parameters.

This approach is similar to the one used in Rost et al. (1997), Eid and Rauber (2000), and Austin et al. (2006) in that mixed Rasch models are applied to identify latent classes that differ in their response style and that the response styles are identified using threshold parameters. However, our approach goes beyond these studies in ensuring that the same trait is being measured across classes. One advantage of our approach is that it is model-based. Other model-based approaches for exploring response styles include using a Bayesian hierarchical model (Rossi, Gilula, & Allenby, 2001), as demonstrated by Buckley (2009), or incorporating response style as a unique dimension in a multidimensional model (Bolt & Johnson, 2009). Another advantage of the approach using the mixed PCM and the constrained mixed PCM is that - compared to indices based on the frequency of extreme or non-extreme responses - it is not necessary to use other items than the ones in the questionnaire and to draw a sample of items that are heterogeneous in content. Note that in our approach the scales for which the mixed PCM shows a better fit, indicating trait heterogeneity, cannot be used in subsequent analyses focused on response styles. In this case differences between classes cannot be attributed to response styles alone, but instead it has to be assumed that the different classes of respondents differ in their substantive interpretation of the items. This would require a more general analysis of the items and the differences between the two latent populations. In addition to a thorough investigation of the item content, searching for predictors of class membership or external criteria that are differentially predicted by the trait values from the classes might also shed light on the substantive differences between the classes.

However, with other methods of investigating response styles, different effects such as the heterogeneity of participants regarding the trait or regarding response scale use may be confounded. In this case results cannot be ascribed to response styles with certainty. For example, if the goal is to analyze the influence of individual differences in response styles on trait scores and participants in different response style groups also show trait heterogeneity, conclusions drawn will be invalid. Thus, it seems preferable to only use scales that allow clear inferences to be drawn concerning response styles. One limitation is the sequential procedure applied in this paper. To conduct both steps, namely the identification of response style groups and the analysis of their consistency across traits, simultaneously in one model was not feasible here due to software restrictions but would be an interesting avenue for future research.

4.2. Implications of the occurrence and consistency of response styles

In other studies investigating response styles using mixed Rasch models (e.g., Rost et al., 1997) a two-class solution consisting of one class of extreme responders and one class of non-extreme responders usually fit the data best. In our study, two latent classes were adequate to describe the participants' response behavior on the PISA 2006 attitude scales and on most NEO-PI-R facets as well. These two latent classes show systematic differences in their endorsement probability for the extreme response categories. For openness to actions and deliberation, a third class emerged in addition to NERS and ERS which was also characterized by an avoidance of extreme categories but differed from the other NERS class in its use of the middle category. Thus, NERS and ERS occur across scales within an instrument and across instruments. The model comparison between mixed PCM and constrained mixed PCM revealed that both instruments contained scales which were not unidimensional, i.e., participants were not homogeneous regarding the trait being assessed. Unidimensional models assume that the endorsement probability for an item is the same for all participants of the same trait level; this is not the case if participants differ in their response style. Thus, trait scores may be distorted by response styles. This was shown by contrasting trait correlations based on a standard PCM which does not take response styles into account with trait correlations based on a constrained mixed PCM in which trait values are estimated separately for each response style class. Standard PCM-based trait correlations were higher than corrected constrained mixed PCM-based trait correlations for almost all scales. The second order LCA illustrated the reason for this result: The response styles were stable for a large part of both samples, demonstrating that respondents using the extreme categories on one scale also tended to do so on other scales. This effect introduced systematic variation across scales that raised the correlations. These results imply that not taking response styles into account can increase trait correlations and that correlations may be over-estimated due to response style effects. When response styles are taken into account, correlations between trait values are lower. Furthermore, while standard PCM trait values and corrected constrained mixed PCM trait values correlated highly with each other, none of the correlations was above .95, indicating that trait values differ, depending on whether response styles are taken into account or not.

The distortion of trait scores by response styles can be countered by using trait values derived from the constrained mixed PCM, since it adjusts latent trait values for response styles (see also Rost et al., 1997). For NERS, trait values are extended whereas for ERS, trait values are contracted compared to the raw sum scores. For example, for two persons with equal sum scores (e.g., 20 on the neuroticism facet anxiety) but different response styles, the person employing NERS will receive a higher trait value (weighted likelihood estimate of 0.69) while the person employing ERS will receive a lower trait value (weighted likelihood estimate of 0.34). Thus, while the raw sum scores for participants using different response styles cannot be compared, the latent trait values resulting from the constrained mixed PCM are comparable. It follows that when response styles have an effect on the data it is preferable to use trait values than sum scores for further analyses.

The second order latent class analysis showed that the two main classes of extreme response style and non-extreme response style occurred consistently across traits for the majority of the participants. That is, between about 65% and 80% of the participants

applied the same response style on every scale in the two instruments we investigated. These participants appear to generally prefer extreme or middle categories, independently of the trait being assessed. The other 20–35% of the participants could not be classified unambiguously as one response style. Results concerning the consistency of NERS coincided well between the analyses on the PISA sample and the NEO sample for a large percentage of participants. For the NEO sample, 5–7% of the participants consistently showed an ERS. This group of participants presumably also exists in the PISA sample, but here was included in the second class which contained participants who by tendency used ERS consistently and participants who switched between response styles. Possible reasons why the consistent ERS class did not emerge in the PISA sample include that the PISA sample was more homogeneous (school children between 15 and 16 years) compared to the NEO sample (age range from 16 to 91). Furthermore, the response format differed between the instruments; the PISA rating scale contained four response categories while the NEO-PI-R additionally contained the middle category *neutral*. Systematic differences in the use of the middle category (Hernández et al., 2004) may have facilitated the occurrence of more latent classes in the NEO sample.

In both the NEO sample and the PISA sample several latent classes were derived, indicating that qualitative differences exist between subgroups of participants in the consistency of their response style. These latent classes contain differing proportions of extreme responders and non-extreme responders across the scales. Descriptively, however, the allocation of the latent classes to NERS or ERS across the scales appears to differ mainly in level. This indicates that response styles might also adequately be modeled as continuous variables (for an example see Bolt & Johnson, 2009). These quantitative differences between response style groups and the consistency of response styles across scales for the majority of the participants raise the question of whether response styles could be modeled as a latent trait variable. If the response style were consistent across traits, it could be modeled using the same response style dimension across different scales. This would also further distinguish response styles from other response tendencies such as faking, which is characterized by intentional behavior (MacCann, Ziegler, & Roberts, 2012) and only occurs on scales which the respondents judge to be relevant to, for example, the job they are applying for (Birkeland, Manson, Kisamore, Brannick, & Smith, 2006). However, a quarter to a third of the participants switched between response styles and response styles have also been shown to vary depending on the response format (Weijters, Cabootor et al., 2010). Further factors that may influence the consistency of response styles are the similarity of the scales regarding content and the social desirability of the items. Thus, the tendency to use response styles may be based on an individual disposition, but the extent to which response styles occur consistently appears to depend on situational factors. Further research could aim at elucidating both trait and situational factors associated with the consistency of response styles.

Acknowledgments

The authors would like to express their gratitude to Dr. Fritz Ostendorf for providing the data for the standardization sample of the German NEO-PI-R. The authors further wish to thank Prof. Dr. Matthias Ziegler for his valuable comments on an earlier version of this paper.

References

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika, 38*(1), 123–140. http://dx.doi.org/10.1007/BF02291180.

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. http://dx.doi.org/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. http://dx.doi.org/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*(2), 143–156. http://dx.doi.org/10.1509/jmkr.38.2.143.18840.

Birkeland, S. A., Manson, T. M., Kisamore, J. L., Brannick, M. T., & Smith, M. A. (2006). A meta-analytic investigation of job applicant faking on personality measures. *International Journal of Selection and Assessment, 14*(4), 317–335. http://dx.doi.org/10.1111/j.1468-2389.2006.00354.x.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*(4), 443–459. http://dx.doi.org/10.1007/BF02293801.

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. http://dx.doi.org/10.1177/0146621608329891.

Borkenau, P., & Ostendorf, F. (1993). *NEO-Fünf-Faktoren-Inventar (NEO-FFI) nach Costa und McCrae*. Göttingen, Germany: Hogrefe.

Bozdogan, H. (1987). Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52*(3), 345–370. http://dx.doi.org/10.1007/BF02294361.

Buckley, J. (2009). *Cross-national response styles in international educational assessments: Evidence from PISA 2006* <https://www.edsurveys.rti.org/PISA/> Retrieved January 2012.

Cattell, R. B., Cattell, A. K. S., & Cattell, H. E. P. (1993). *16PF fifth edition questionnaire*. Champaign, IL: Institute for Personality and Ability.

Costa, P. T., & McCrae, R. R. (1992). *Revised NEO personality inventory (NEO-PI-R) and NEO Five-Factor inventory (NEO-FFI)*. Odessa, FL: Psychological Assessment Resources.

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

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Frey, A., Taskinen, P., Schütte, K., Prenzel, M., Artelt, C., Baumert, J., et al. (2009). *PISA 2006 Skalenhandbuch: Dokumentation der Erhebungsinstrumente* [PISA 2006 Scales handbook: Documentation of the instruments]. Münster, Germany: Waxmann.

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. http://dx.doi.org/10.1037/0021-9010.89.4.687.

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.

Keller, F., & Kempf, W. (1997). Some latent trait and latent class analyses of the Beck-Depression-Inventory (BDI). In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 314–323). Münster, Germany: Waxmann.

MacCann, C., Ziegler, M., & Roberts, R. D. (2012). Faking in personality assessment: Reflections and recommendations. In M. Ziegler, C. MacCann, & R. D. Roberts (Eds.), *New perspectives on faking in personality assessment* (pp. 309–329). New York, NY: Oxford University Press.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149–174. http://dx.doi.org/10.1007/BF02296272.

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

OECD (2006). *PISA 2006 assessment framework*. Paris, France: OECD Publications.

Ostendorf, F., & Angleitner, A. (2004). *NEO-PI-R: NEO-Persönlichkeitsinventar nach Costa und McCrae*. Göttingen, Germany: Hogrefe.

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: Academic Press.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen, Denmark: Danmarks Paedagogiske Institut.

Rossi, P. E., Gilula, Z., & Allenby, G. M. (2001). Overcoming scale usage heterogeneity: A Bayesian hierarchical approach. *Journal of the American Statistical Association, 96*(1), 20–31. http://dx.doi.org/10.1198/016214501750332668.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement, 14*(3), 271–282. http://dx.doi.org/10.1177/014662169001400305.

Rost, J. (1991). A logistic mixture distribution model for polytomous item responses. *The British Journal for Mathematical and Statistical Psychology, 44*, 75–92.

Rost, J. (2004). *Lehrbuch Testtheorie – Testkonstruktion* [Textbook test theory – test construction] (2nd ed.). Bern, Switzerland: Verlag Hans Huber.

Rost, J., Carstensen, C. H., & von Davier, M. (1997). Applying the Mixed Rasch Model to personality questionnaires. In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 324–332). Münster, Germany: Waxmann.

Smith, P. B. (2004). Acquiescent response bias as an aspect of cultural communication style. *Journal of Cross-Cultural Psychology, 35*(1), 50–61. http://dx.doi.org/10.1177/0022022103260380.

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. http://dx.doi.org/10.1177/0022022104264126.

von Davier, M.(2001). *WINMIRA 2001* [Computer software]. Kiel, Germany: Institute for Science Education.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*(3), 427–450. http://dx.doi.org/10.1007/BF02294627.

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. http://dx.doi.org/10.1016/j.ijresmar.2010.02.004.

Weijters, B., Geuens, M., & Schillewaert, N. (2010a). The individual consistency of acquiescence and extreme response style in self-report questionnaires. *Applied Psychological Measurement, 34*(2), 105–121. http://dx.doi.org/10.1177/0146621609338593.

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

Wetzel, E., Böhnke, J. R., Carstensen, C. H., Ziegler, M., & Ostendorf, F. (2012) Do response styles matter? Assessing differential item functioning for men and women in the NEO-PI-R. Submitted for pulication.

2.3. Appendix C: Manuscript Wetzel & Carstensen (2013a)

Linking PISA 2000 and PISA 2009:

Implications of instrument design on measurement invariance

Abstract

An important pre-requisite of trend analyses in large scale educational assessments is the measurement invariance of the testing instruments across cycles. This paper investigates the measurement invariance of the PISA 2000 and PISA 2009 reading instruments using Item Response Theory models. Links between the PISA 2000 and PISA 2009 instruments were analyzed using data from a sample tested in 2009 which took both the PISA 2000 and PISA 2009 instruments and additionally using part of the German PISA 2000 sample as well. Model fit comparisons showed that the instruments are not measurement invariant and that some link items show large differences in item difficulty. Position effects may explain some of these differences and may also influence the size of the link error.

Key words: PISA, measurement invariance, linking, link error, position effects

Introduction

To introduce the aim of the present paper, we will give a brief introduction to the goals and study design of the Programme for International Student Assessment (PISA). Second, we will describe the linking of scores from different PISA assessments and introduce the computation of the link error, and third, we will present the aims of our study and our research questions.

Goal and Study Design of PISA

Starting in the year 2000, the Organisation for Economic Cooperation and Development (OECD) has been conducting the Programme for International Student Assessment (PISA) which assesses 15-year-olds every three years in the domains of reading, mathematics, and science. The aim of PISA is to measure life skills that enable people to succeed in modern societies (e.g., OECD, 2009a). Accordingly, PISA requires students to evaluate material and apply it to new situations. The three domains are defined in terms of a literacy concept similar to the one developed by previous surveys, for example the International Adult Literacy Survey (IALS; e.g., OECD & Statistics Canada, 2000). Reading literacy is characterized by a person's capacity to "understand, use, reflect on and engage with written texts, in order to achieve one's goals, to develop one's knowledge and potential, and to participate in society" (OECD, 2009a; p. 14). Mathematical literacy is defined as "an individual's capacity to identify and understand the role that mathematics plays in the world, to make well-founded judgements and to use and engage with mathematics in ways that meet the needs of that individual's life as a constructive, concerned and reflective citizen." (OECD, 2009a; p. 14). Scientific literacy comprises "an individual's scientific knowledge and use of that knowledge to identify questions, to acquire new knowledge, to explain scientific phenomena, and to draw evidence-based conclusions about science-related issues, understanding of the characteristic features of science as a form

of human knowledge and enquiry, awareness of how science and technology shape our material, intellectual, and cultural environments, and willingness to engage in science-related issues, and with the ideas of science, as a reflective citizen." (OECD, 2009a; p. 14).

In PISA, the main focus of the study alternates. In 2000 it was reading, in 2003 mathematics, and in 2006 science. With the completion of the fourth PISA assessment in 2009, a new cycle has begun in which reading was once again the first major domain. The major domain is assigned more testing time than the minor domains. In general, items are nested in units (e.g., items that refer to the same text passage) and several units compose a cluster. The items in one cluster all assess the same domain. Each test booklet contains four clusters. The test booklets are randomly assigned to the students participating in PISA. Comparisons of student achievements in the three domains across the participating countries have been drawn from the first PISA study in 2000 and continue to give important information regarding the standing of students in one nation compared to others. Another central goal of PISA which is increasingly taking priority is conducting trend analyses. Trend analyses aim at investigating how student achievements develop within participating countries over assessment periods (OECD, 2010). Trend analyses (with regard to the whole population or subpopulations) carry critical implications as they can be used to monitor the success of reforms in educational systems. For instance, policy makers may be interested in whether the proportion of low-achieving students has decreased or whether the potential gender gap in achievement has narrowed or widened.

Linking and the Link Error

Conducting methodologically sound trend analyses is not an easy task. One pre-requisite for trend analyses is the measurement invariance of the instruments across assessments (Kolen & Brennan, 2004). In their review of the PISA test design, Mazzeo and von Davier

(2009) list several criteria that need to be fulfilled to establish stable trends. These include that the same construct should be measured in all assessments and in all participating countries. Furthermore, the relationship between the items and the underlying latent trait should be unchanged across assessments for items that are used in several assessments. Also, item presentation should be standardized and comparable across countries and assessments.

To ensure the comparability of scores from different assessments, link items, which are common across assessments, are used. For example, 28 of the 129 reading items used in PISA 2000 were included in PISA 2003, 2006, and 2009. Changes in the difficulty of these link items determine the transformation used to equate scores from one assessment with scores from a previous assessment (OECD, 2012). Since the chosen link items are a sample of all possible link items, a different transformation would result if an alternative set of link items had been chosen. Thus, uncertainty is introduced to the process of equating scores across data collections. The precision with which scores from different assessments are aligned on one performance scale is captured by the link error (or equating error). The computation of the PISA 2003 link error was shown to be inadequate by Monseur and Berezner (2007), so it was modified to take into account that items are organized in units and that partial credit items have a greater influence on scores than dichotomous items. The improved link error estimate has been used to link PISA 2009 and PISA 2006 data to previous data collections and is described in the PISA technical reports (OECD 2009b, 2012). First, the difference in item difficulty $\hat{\delta}_{ij}$ between two assessments (e.g., PISA 2009 and PISA 2006) is computed $c_{ij} = \hat{\delta}_{ij}^{2009} - \hat{\delta}_{ij}^{2006}$ with *i* items in a unit and $j = 1,...,K$ units. The mean number of score points is $\bar{m} = \frac{1}{K}\sum_{j=1}^{K} m_j$. Further it is defined that $c_{\bullet j} = \frac{1}{m_j}\sum_{j=1}^{m_j} c_{ij}$ and $\bar{c} = \frac{1}{N}\sum_{j=1}^{K}\sum_{i=1}^{m_j} c_{ij}$. Then the link error can be computed as

$$error_{2009,2006} = \sqrt{\frac{\sum_{j=1}^{K} m_j^2 (c_{\bullet j} - \bar{c})^2}{K(K-1)\bar{m}^2}} \,. \qquad (1)$$

PISA reported the link error to be 4.07 for the reading scale 2006 to 2009 and 4.94 for the reading scale 2000 to 2009 (OECD, 2012). Thus, when taking only the link error into account, the 95% confidence interval of the difference in mean scores is about 20 score points wide (Wu, 2010). Monseur and Berezner (2007) also argued that the link error may be larger than the sampling error and the measurement error. The link error influences trend results and conclusions drawn from trend analyses and as such has an effect on actions taken by policy-makers. Gebhardt and Adams (2007) demonstrated that trend results differed depending on whether international item parameters were used or whether national item para-meters were used in computation. Since link errors threaten trend analyses, both Mazzeo and von Davier (2009) and Wu (2010) recommend increasing the number of link items to reduce the link error.

As linking is such an important aspect of trend analyses, this study investigates the linking of PISA 2000 and PISA 2009 reading and science items for two German samples. In 2009, the German PISA consortium conducted an additional study to the regular PISA 2009 assessment in which the PISA 2009 booklets as well as five selected booklets from the PISA 2000 assessment were administered to students at 59 German high schools. These 59 high schools had already participated in PISA 2000-E as part of an extended sample for state comparisons (Baumert et al., 2002). Thus, data were available from the same 59 high schools for two different time points, 2000 and 2009, as well as items from two different PISA instruments, namely the PISA 2000 and the PISA 2009 test booklets. This design allowed the measurement invariance of the PISA 2000 and PISA 2009 reading instruments to be investigated within one sample (the sample from 2009) as well as between samples within one instrument (PISA 2000). The five booklets from PISA 2000 applied in 2009 originally contained mathematics and science items at the last cluster position. The last clusters in these five booklets

were replaced with science clusters from the PISA 2006 assessment, enabling us to analyze the measurement invariance of the PISA 2006 and PISA 2009 science instruments for 44 out of 53 science link items as well.

The aim of this paper is to test the measurement invariance of the reading items from 2000 and 2009 regarding the common items and link items and the science items from 2006 and 2009 regarding a subset of the link items. Our goal is to examine whether it is possible to establish a link and if so, which items are adequate for establishing a stable link. Furthermore, trend results will be reported and factors that influence linkability will be discussed in terms of how they affect the size of the link error. One conceivable influence on linking are position effects, i.e., the phenomenon that items have different difficulties, depending on their position in the test. For PISA 2000, Adams and Carstensen (2002) showed that differences in item difficulties between positions occurred for each of the nine reading clusters. Position effects are possible in PISA because clusters contain different units of items between assessments, as some items are replaced and as changes in testing time need to be accommodated when the major domain alternates. Thus, it will be analyzed whether differences in position may account for differences in item difficulties across assessments and instruments.

The samples used in this study allow the assessment of measurement invariance from two perspectives, first concerning the link and common items in the PISA 2000 and PISA 2009 instruments and second concerning the link and common items in the PISA 2000 reading instrument for which data was collected in 2000 and 2009. Thus, in sum, our two main research questions are 1) whether the instruments from PISA 2000 and PISA 2009 are invariant regarding the reading link and common items and whether the instruments from PISA 2006 and PISA 2009 are invariant regarding the science link items for the same study undertaken in 2009 and 2) whether the instrument from PISA 2000 is invariant between different studies (2000 vs. 2009) regarding the reading link and common items.

Method

Instrument

Reading clusters from PISA 2000 and PISA 2009 as well as science clusters from PISA 2006 and PISA 2009 were used. Table 1 lists the number of items linking the assessments. As the number of common reading items between 2000 and 2009 (39 items) is larger than the number of link items (28), analyses will be conducted (a) with the common items and (b) with the link items. As only items being used repeatedly between assessments were analyzed, subscales for the different domains were not taken into account. A list of all the items included in our analyses as well as the item parameter estimates obtained from separate partial credit models in each of the subsamples can be found in the Appendix.

Sample

Two datasets were combined to obtain the dataset analyzed here. Both datasets were collected from 9th graders at the same 59 German high schools, though during different assessments. The first dataset ("study 2000") consisted of 1487 students (54.2 % female) who were regular participants of the PISA 2000-E (Baumert et al., 2002) assessment in Germany. The booklet design of the PISA 2000 study is depicted in Table 2. The second sample ("study 2009"; $N = 1948$, 53.6% female) formed an additional sample to the German PISA 2009 sample. For this second sample, both the 13 new PISA 2009 booklets (with regular difficulty; OECD, 2012) as well as five additional booklets (OECD, 2002) were applied (see Table 3). These 18 booklets were randomly distributed, resulting in a subsample of 1394 students who filled out the PISA 2009 booklets (booklets 1 - 13) and a subsample of 554 students who filled out booklets 14 to 18. Booklets 14 to 18 contained reading clusters from PISA 2000 at cluster positions one to three, regarding these three clusters they were identical to booklets 1 to 5 in the original PISA 2000 assessment (see Tables 2 and 3). The last cluster in the PISA 2000

booklets was originally used for mathematics and science items; for our study this cluster was replaced by a science cluster from the PISA 2006 assessment. To differentiate between the different item sets, each instrument will be referred to by its domain (reading or science) and PISA study year that the items originated from, e.g., "reading 2000" refers to the reading items from the PISA 2000 instrument. Thus, booklets 14 to 18 are a combination of reading 2000 (cluster positions 1 – 3) and science 2006 (cluster position 4).

Table 1

PISA Link Items across Assessments for the Three Domains

		Instrument			
		PISA 2000	PISA 2003	PISA 2006	PISA 2009
Domain	Reading	**129 items**	28 link items 00/03/06	28 link items 00/03/06	**39 common items 00/09, 28 link items 00/03/06/09**
	Mathematics	20 link items 00/03	**84 items**	48 link items 03/06	35 link items 03/06/09
	Science	25 link items 00/03	22 link items 03/06	**108 items**	53 link items 06/09

Note. Major domains are depicted in boldface and the absolute number of items is reported.

Analyses

Measurement invariance was assessed from two perspectives. The first research question asked whether the instruments from PISA 2000 and PISA 2009 were measurement invariant regarding the reading items from the same study in 2009. For the science items, this question pertained to the instruments from PISA 2006 and PISA 2009. The second research question was whether the instrument from PISA 2000 was measurement invariant for different studies (study 2000 vs. study 2009). This question was analyzed using the reading items.

To answer these research questions, random coefficients multinomial logit models (RCMLM; Adams & Wilson, 1996) were estimated using ConQuest (Wu, Adams, Wilson, &

Haldane, 2007). The RCMLM is a flexible generalization of the Rasch model (RM; Rasch, 1960) which integrates other Rasch-type models such as the rating scale model (Andrich, 1978), the partial credit model (PCM; Masters, 1982), multifaceted models (Linacre, 1994), and the linear logistic test model (LLTM; Fischer, 1973). Thus, the RCMLM allows group differences (e.g., between study 2000 and study 2009) to be incorporated into the model as well as item by group interactions (differential item functioning).

The basic model for items with dichotomous response formats was the Rasch model (Rasch, 1960)[3] which models the probability that person *v* with person parameter θ_v will give a correct response to item *i* with difficulty δ_i :

$$p(X_{vi} = 1 \mid \theta_v, \delta_i) = \frac{\exp(\theta_v - \delta_i)}{1 + \exp(\theta_v - \delta_i)}. \quad (2)$$

Equation 2 can also be expressed in logit form:

$$\text{logit} = \ln \frac{p(X_{vi} = 1)}{1 - p(X_{vi} = 1)} = \theta_v - \delta_i. \quad (3)$$

The item difficulty δ_i can further be parameterized to account for properties that certain items share (e.g., cognitive operations involved in solving them). In this case, the LLTM (Fischer, 1973) results: $\text{logit} = \theta_v - \sum_{k=0}^{K} \eta_k \omega_{ik}$ where η_k is a difficulty parameter for item property *k* and ω_{ik} represents an indicator weight of item *i* on item property *k* which takes the value 1 if item *i* belongs to property *k* and 0 otherwise. Two extensions of this model were compared to test measurement invariance. Model 1 consisted of an RM and a unique mean parameter β_g with *g = 1, ..., G* for the student performance distribution in the respective study or instrument:

[3] For partial credit items the partial credit model (Masters, 1982) was used.

Model 1: $$\text{logit} = \theta_v - \delta_i + \beta_g. \quad (4)$$

Model 2 additionally modeled the interaction between study or instrument and the difficulty of the item:

Model 2: $$\text{logit} = \theta_v - \delta_{ig} + \beta_g. \quad (5)$$

That is, in Model 2, differences in item difficulties (differential item functioning) between the studies or instruments were also estimated[4]. To evaluate the magnitude of these differences, the classification system for differential item functioning (DIF) developed by Educational Testing Service (ETS) was applied. In this classification system, items with DIF values below .25 contain negligible DIF, items with DIF values between .25 and .37 contain slight to moderate DIF, and items with DIF values equal to or above .38 contain moderate to large DIF (Zieky, 1993). For reading, the two models were computed once with the link items and a second time using the common items.

ConQuest applies marginal maximum likelihood estimation using an EM algorithm (Bock & Aitkin, 1981) to estimate the item parameters and a normally distributed ability density. For the model comparisons, the mean of the item parameters was constrained to be zero for model identification purposes. Note that for the PCMs reported in the Appendix the model identification constraint was placed on the cases, yielding a mean latent variable of zero. Missing values were treated according to the PISA procedure (e.g., OECD, 2012). That is, responses to items that the student had reached and were missing or invalid were recoded as incorrect while items that the student had not reached were treated as not administered. Comparisons of model fit between the models test the assumption that differences in item difficulties between assessments are negligible and that joint scaling can therefore be conducted

[4] ConQuest model statements for the two models are :
Model 1: item + item*step + instrument
Model 2: item + item*step + instrument + item*instrument

across assessment periods. The difference in the deviance (-2 x log-likelihood) of the two models was tested for significance using a χ^2-test. Furthermore, Akaike's Information Criterion (AIC; Akaike, 1973), the Bayesian Information Criterion (BIC; Schwarz, 1978), and the consistent Akaike's Information Criterion (CAIC; Bozdogan, 1987) were consulted for finding the better-fitting model. Note that standard errors reported for estimates on group differences and estimates of the interaction between study or instrument and the item do not take into account the link error and neither the sampling error but only represent the statistical uncertainty due to parameter estimation.

Furthermore, the link error (see introduction) was investigated. The link error was computed for the reading link items, the common reading items, and the science link items for each of the different combinations between study and instrument. The link errors for the common reading items and the science link items were compared to the ones reported in the PISA 2009 Technical Report (OECD, 2012). Additionally, the link error 2000/2009 was computed separately by cluster position for the common reading items to investigate whether there were differences in the size of the link error depending on which position in the test booklet the items were located. As the common reading items were only on positions one to three in PISA 2000 (see Table 2), a separate link error for cluster position four could not be computed. This analysis was conducted using data from the 28 OECD countries that had taken part in PISA 2000 and PISA 2009[5]. For each of the cluster positions, a random sample of about 500 students per country was drawn. Selection probabilities in the random sample should be equal to those in the complete sample. To achieve this, we multiplied the final student weights (which reflect the variation in selection probabilities) with random numbers from a uniform distribution to draw the random sample. For the computation of the link error across all cluster positions,

[5] Public use data from the PISA assessments is available online at http://www.oecd.org/pages/0,3417,en_32252351_32235731_1_1_1_1_1,00.html

both a random sample of about 500 students per country as well as a random sample of about 2000 students per country was drawn.

We analyzed whether differences in the position of items might be explanative for differences in the difficulty between studies and instruments. The PISA test design (from PISA 2003 on) has been balanced regarding the item clusters; that is, each cluster as a whole appears at each of the four cluster positions in one of the test booklets. However, the position of each item unit within its respective cluster is fixed. Thus, the test design is not balanced regarding the position of the item units within clusters. Between assessments the allocation of item units to clusters can change e.g., due to differing amounts of testing time. In consequence, it is possible for an item to differ in its position in the cluster between two PISA assessments. For example, item R055Q01 was at position 9 in cluster R2 in PISA 2009 while it was at position 3 in cluster R5 in PISA 2000. This means that item difficulties are by design always confounded with the position of items in clusters and the positions are not perfectly controlled for due to constraints in test assembly.

To test directly using a model-based approach whether there is an interaction between item position and instrument would be an interesting prospect. However, this would require a balanced design with regard to the item position which provides data for each possible combination of an item with a position in the instrument. Since this is not the case in the design of the presented national add-on study, this model cannot be estimated. To approximate this model we instead extended Model 1 (Equation 4) to include an interaction between item, cluster, and instrument. To test whether there was a meaningful interaction between these three components, we compared Model 1 to a model including this three-way interaction[6] where the cluster position is $c = 1, \ldots, C$:

$$\text{Model 3:} \qquad \text{logit} = \theta_v - \delta_{igc} + \beta_g. \tag{6}$$

[6] item*cluster*instrument in the ConQuest model statement

From PISA 2003 on the test design has been balanced regarding the cluster positions. However, in PISA 2000 this was not yet the case so the model comparison concerning instrument 2000 and instrument 2009 had to be conducted with the set of items that were positioned at all cluster positions in instrument 2000 (17 of the common reading items).

To further test whether position effects may have been responsible for differences in item difficulty, correlations were computed. An index for the cluster position was created which took into account the number of items in each respective cluster, the position of the item within the cluster, and the position of the cluster in the respective booklet. The first value of the index identified the item's cluster position (1, 2, or 3). The fraction consisted of the position of the item within the cluster (counting from 0) divided by the number of items in the cluster: $Index = cluster\ position + \frac{(item\ number - 1)}{N\ items\ in\ cluster}$. For example, item R055Q01 was the ninth of 15 items in reading cluster R2 which was at position 1 in booklet 8. Thus, item R055Q01 received the index 1 (9-1)/15. In booklet 13, cluster R2 was at position 2 and item R055Q01 therefore received the index 2 8/15. Then, these indices were averaged to obtain the mean position of the items (see Appendix 1). The differences in this position index between instruments were correlated with the differences in item difficulty. If the mean position of items differs between instruments, potentially a bias in the item difficulties might be introduced which corresponds to the average position of the items. To quantify this potential bias, differences in item difficulty were regressed on differences in the position index. The potential bias then equals the predicted value in the item difficulty difference for the average difference in position.

Table 2

PISA 2000 Booklet Design

Booklet ID	Cluster 1	2	3	4
1	R1	R2	R4	M1 M2
2	R2	R3	R5	S1 S2
3	R3	R4	R6	M3 M4
4	R4	R5	R7	S3 S4
5	R5	R6	R1	M2 M3
6	R6	R7	R2	S2 S3
7	R7	R1	R3	R8
8	M4 M2	S1 S3	R8	R9
9	S4 S2	M1 M3	R9	R8

Note. R = reading, M = mathematics, S = science.

Table 3

Study 2009 Booklet Design

	Booklet ID	Cluster 1	2	3	4
PISA 2009 booklets	1	M1	R1	R3A	M3
	2	R1	S1	R4A	R7
	3	S1	R3A	M2	S3
	4	R3A	R4A	S2	R2
	5	R4A	M2	R5	M1
	6	R5	R6	R7	R3A
	7	R6	M3	S3	R4A
	8	R2	M1	S1	R6
	9	M2	S2	R6	R1
	10	S2	R5	M3	S1
	11	M3	R7	R2	M2
	12	R7	S3	M1	S2
	13	S3	R2	R1	R5
Reading 2000, Science 2006	14	R1	R2	R4	S1-MS06
	15	R2	R3	R5	S4-MS06
	16	R3	R4	R6	S5-MS06
	17	R4	R5	R7	S6-MS06
	18	R5	R6	R1	S7-MS06

Note. R = reading, M = mathematics, S = science, MS = main study. Booklets 14 to 18 contain reading clusters from PISA 2000 and science clusters from PISA 2006. For some clusters there were two versions, a regular one (A) and an easier one (B).

Results

In the following, results will be reported for the analyses on measurement invariance and regarding the influence of the position. The first section contains the results for reading and the second section contains the results for science.

Reading

Two of the reading items, R219Q01T and R219Q01E, were removed on the international level due to data entry errors as described in the PISA 2009 technical report (OECD, 2012). Thus, the reading link consisted of 26 items and there were 37 common reading items between reading 2000 and reading 2009. The group parameters included in Model 1 show for which subsample the items, taken as a whole, were easier. For study 2009, participants filling out reading 2009 were slightly better compared to participants filling out reading 2000 (-0.02 logits, SE = 0.03) regarding the 37 common items. Concerning the 26 link items, the difference in the group parameter was -0.08 logits (SE = 0.03) for study 2009, again favoring participants tested with the PISA 2009 instrument. For the PISA 2000 instrument, the 37 reading items were easier for participants tested in 2000 compared to participants tested in 2009 (-0.33 logits, SE = 0.03). The 26 reading link items were -0.33 logits (SE = 0.03) easier for students assessed with the PISA 2000 instrument in 2000 compared to students assessed with the same instrument in 2009. Bischof, Hochweber, Hartig, and Klieme (in press) did not find significant differences in the mean reading performance between 2000 and 2009 for samples from the same 59 schools used in our study.

Comparisons of model fit showed that for both item sets and for both research questions, Model 1 had lower BIC and CAIC values compared to Model 2 (see Table 4a and Table 4b). However, both the significant χ^2-tests of the difference in deviance between the models

and the AIC indicated that the more complex Model 2 fit better than Model 1. Thus, meaningful differences in item difficulty appear to exist. A closer investigation of the item difficulties revealed substantial differences for some items. Figure 1a shows the differences in item difficulty for study 2009 between the PISA 2000 instrument and the PISA 2009 instrument. Positive values indicate that the item was more difficult in the PISA 2009 instrument. Most reading items fall in the negligible or slight to moderate category of the ETS classification system (Zieky, 1993), but some clearly exceed the limit for moderate DIF, most notably R055Q03 which was extremely easy for students filling out reading 2009 compared to students filling out reading 2000 (-1.71 logits, SE = 0.09, 95% CI [-1.88, -1.53]). In Figure 1b, the differences in item difficulty for the reading items in the PISA 2000 instrument between study 2000 and study 2009 are depicted. Here, some of the same items as in Figure 1a show large differences (e.g., R220Q05 with 1.14 logits, SE = 0.21, 95% CI [0.73, 1.55]), though others showing large differences in Figure 1a only showed smaller differences in Figure 1b (e.g., R055Q03 with -0.66 logits, SE = 0.08, 95% CI [-0.82, -0.51]). The pattern of the item difficulty differences is very similar between the common and link items. However, for some items (e.g., R220Q05) the difference is marginally larger when all 37 common items (1.14 logits, SE = 0.21, 95% CI [0.73, 1.55]) are included compared to only the 26 link items (1.06 logits, SE = 0.21, 95% CI [0.65, 1.46]) in the comparison of study 2000 and study 2009 for the PISA 2000 instrument. For other items (e.g., R219Q02), the difference is slightly smaller with 37 items (-0.45 logits, SE = 0.14, 95% CI [-0.73, -0.17]) compared to 26 items (-0.52 logits, SE = 0.14, 95% CI [-0.80, -0.24]).

Model 1 was recomputed after removal of the items with the largest differences in item difficulty, namely R055Q03 and R220Q05 for the reading link items and additionally R101Q02 for the common reading items. First, group differences were re-assessed for the comparison of the PISA 2000 and PISA 2009 instruments in study 2009. The remaining 24

reading link items did not differ in difficulty between participants tested with reading 2000 in study 2009 compared to participants tested with reading 2009 in the same study (0.00 logits, SE = 0.03). The reduced number of 34 common items yielded a group parameter of -0.04 logits (SE = 0.03) for study 2009 with participants filling out reading 2000 having slightly better results. The group difference for the common items changed its direction and is slightly larger compared to the full item set.

Second, group differences were re-assessed for the comparison of reading 2000 in study 2000 and study 2009. Regarding the remaining 24 link items, the group parameter amounted to -0.37 logits (SE = 0.03), indicating that the PISA 2000 instrument was easier for students in study 2000 compared to students in study 2009. For the 34 common items the group parameter was -0.35 logits (SE = 0.03), indicating that it too was easier for students assessed in study 2000 compared to students assessed in study 2009. Thus, concerning reading 2000 in study 2000 and study 2009, the differences are in the same direction and slightly larger compared to the full item set for both the link items and the common items.

Link errors were computed between the instruments from PISA 2000 and PISA 2009 for the reading link and common items. These are reported in Table 5. For example, for the 37 common reading items, the link error was 6.43 points on the PISA reading scale for the link between the instruments from PISA 2000 and PISA 2009 both applied in study 2009. When only the 34 common reading items without large differences in item difficulty were used, the link error decreased to 5.48 points. The OECD reports a link error of 4.94 on the PISA reading scale between 2000 and 2009 (Table 12.36; OECD, 2011) which is lower than the link errors computed with our data. For comparisons of the magnitude of the link error in relation to the cluster position of the items in the booklets, the link error was computed separately by cluster position for the 37 common reading items using data from an international sample with N = about 500 per OECD country. When the common reading items were at cluster position 1 in

the test booklet, a link error of 6.69 points on the PISA reading scale resulted between PISA 2000 and PISA 2009 (see Table 5). Across the three cluster positions, the international sample with about 500 students per OECD country yielded a link error of 5.86 points while the international sample with about 2000 students per OECD country yielded a link error of 5.92 points on the PISA reading scale.

To investigate one possible reason for the differences in item difficulty we found, position effects were estimated. First, we compared the model fit between Model 1 and Model 3 (including the three-way interaction between item, cluster, and instrument) for the 17 reading items that appeared at all three cluster positions. Model 1 yielded an AIC of 12060.84 (BIC = 12194.63, CAIC = 12218.63) while Model 3 yielded an AIC of 12088.89 (BIC = 12412.22, CAIC = 12470.22). Thus, the simpler model showed a better fit indicating that overall differences in item position did not play an important role for differences in item difficulty between the two instruments.

Second, we investigated correlations between differences in cluster position and differences in item difficulty. The correlation between the difference in cluster position (reading 2009 – reading 2000) and the difference in item difficulty between the two instruments was $r = .29$ ($p = .08$; $N = 37$) for the 37 common reading items. When the three items with large item difficulty differences were not included, the correlation rose to $r = .41$ ($p = .02$). The mean difference in cluster position was -0.19 ($SD = 0.55$) which corresponds to about one fifth of a cluster's length. Thus, common reading items on average were at a slightly earlier position (about three to four items earlier) in the PISA 2009 instrument. The resulting potential bias (quantified as the predicted value for the average difference in item position in the regression of difference in item difficulty on difference in item position) for the 34 remaining common items amounts to -0.05 logits (CI [-0.11, 0.01]) in favor of students tested with the PISA 2009 instrument. For the 26 link items the correlation between the difference in cluster position and

the difference in item difficulty was $r = .26$ ($p = .20$). For the reduced item set of 24 reading link items, this correlation increased to $r = .55$ ($p = .01$). The 24 link items yielded a mean difference in cluster position of -0.29 ($SD = 0.44$) and a potential bias of -0.13 logits (CI [-0.21, -0.05]). Thus, we would expect participants taking reading 2009 to be slightly better compared to participants taking reading 2000 solely based on the earlier position of the reading items for both reading item sets, though the bias is larger when only taking the link items into account. However, as noted above, in our data students taking reading 2009 were only better than students taking reading 2000 for all 37 common item and the 26 link items, but not for the reduced set of 34 common items. Note that confidence intervals were computed taking into account only the regression's standard error. Considering the measurement error, the sampling error, and the link error additionally would result in wider confidence intervals for the potential bias.

Table 4a

Comparison of Model Fit for the PISA 2000 (2006) Instrument and the PISA 2009 Instrument for Reading and Science

Domain	Model	N	#par	-2 ln*L*	AIC	BIC	CAIC	X^2	df	p
Reading (37 common items)	1 PCM + instrument	1948	45	24355.02	24445.02	24695.88	24740.88			
	2 PCM + instrument + instrument*item	1948	81	24171.27	24333.27	24784.80	24865.8	183.76	36	<.001
Reading (26 link items)	1 PCM + instrument	1948	34	18645.13	18713.13	18902.67	18936.67			
	2 PCM + instrument + instrument*item	1948	59	18483.56	18601.56	18930.46	18989.46	161.57	25	<.001
Science	1 PCM + instrument	1948	47	11337.18	11431.18	11693.19	11740.19			
	2 PCM + instrument + instrument*item	1948	90	11176.83	11356.83	11858.54	11948.54	160.36	43	<.001

Note. PCM = partial credit model, #par = number of parameters, L = Likelihood, BIC = Bayesian Information Criterion, CAIC = consistent Akaike's Information Criterion, χ^2 = -2lnL model 1 – (-2lnL model 2), df = #par model 1 - #par model 2.

Table 4b

Comparison of Model Fit for Study 2000 and Study 2009

Domain	Model	N	#par	-2 ln*L*	AIC	BIC	CAIC	X^2	df	p
Reading (37 common items)	1 PCM + study	2041	45	24262.32	24352.32	24605.28	24650.28			
	2 PCM + study + study*item	2041	81	24167.96	24329.96	24785.28	24866.28	94.36	36	<.001
Reading (26 link items)	1 PCM + study	2041	34	18928.30	18996.30	19187.42	19221.42			
	2 PCM + study + study*item	2041	59	18852.76	18970.76	19302.41	19361.41	75.54	25	<.001

Note. PCM = partial credit model, #par = number of parameters, L = Likelihood, BIC = Bayesian Information Criterion, CAIC = consistent Akaike's Information Criterion, χ^2 = -2lnL model 1 – (-2lnL model 2), df = #par model 1 - #par model 2.

Table 5

Link errors for Reading and Science

Domain and sample	**Number of items**			
Reading	**37 common items**	**34 reduced common items**	**26 link items**	**24 reduced link items**
Reading 2000 and 2009 in study 2009	6.43	5.48	6.33	6.30
Reading 2000 in study 2000 and study 2009	6.30	4.34	8.02	5.66
International sample				
Cluster position 1 (N = 500 per country)	6.69			
Cluster position 2 (N = 500 per country)	8.13			
Cluster position 3 (N = 500 per country)	8.13			
Cluster positions 1-3 (N = 2000 per country)	5.92			
Cluster positions 1-3 (N = 500 per country)	5.86			

Science	**44 link items**	**42 reduced link items**
Science 2006 and Science 2009 in study 2009	10.50	7.55

Note. Link errors for reading are reported on the PISA 2000 scale. Link errors for science are reported on the PISA 2006 scale.

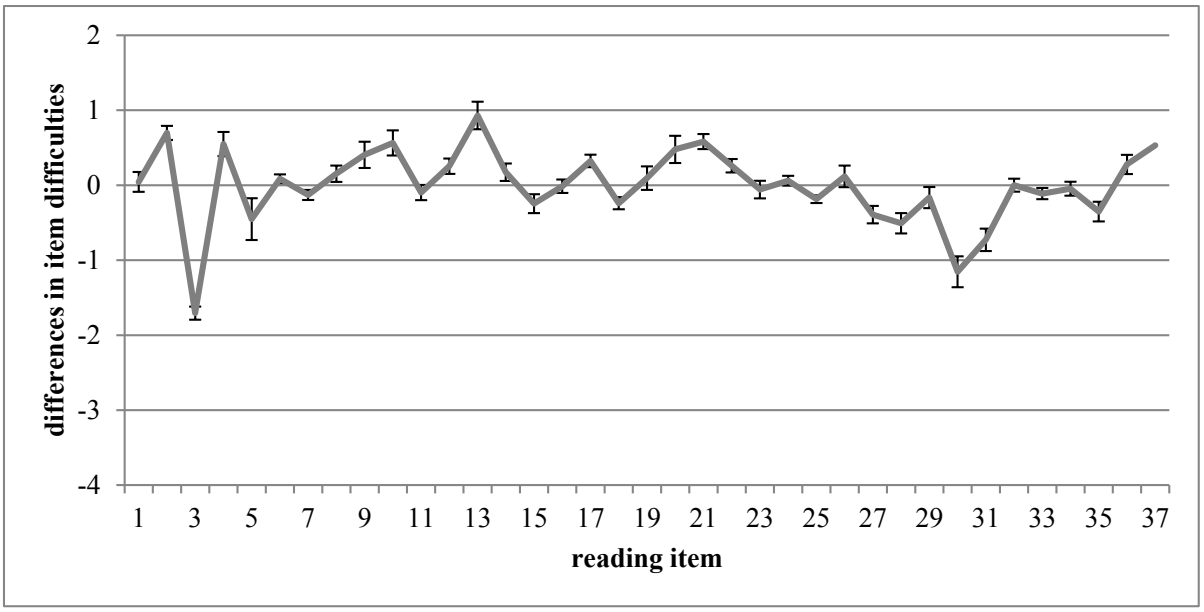

Figure 1a. Differences in item difficulties for reading between the PISA 2009 instrument and the PISA 2000 instrument in study 2009. Positive values indicate that the item was more difficult in the PISA 2009 instrument. Error bars represent standard errors.

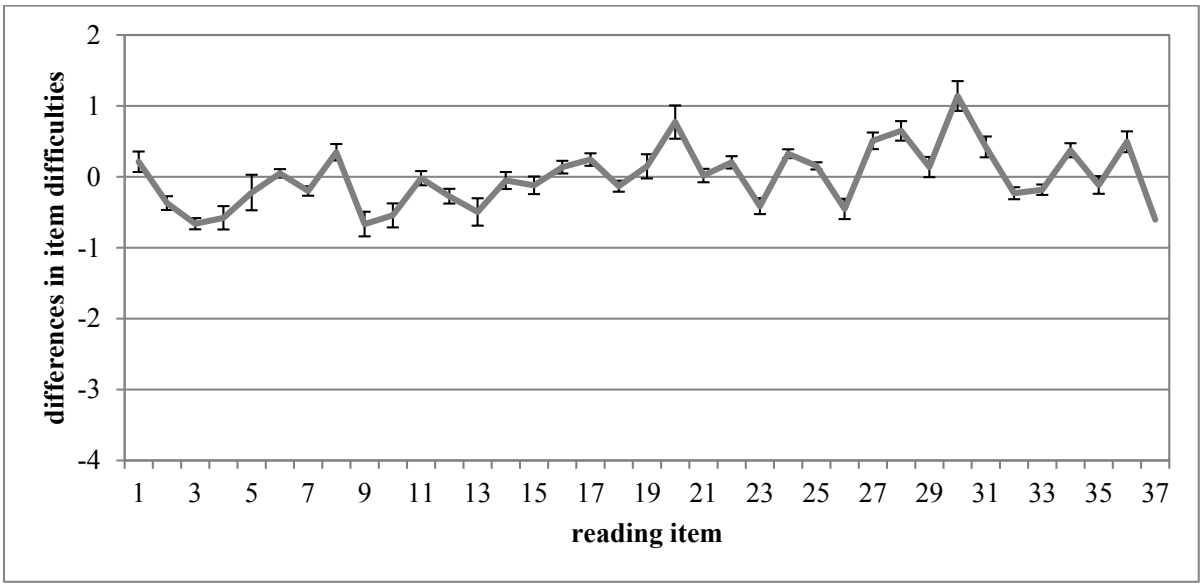

Figure 1b. Differences in item difficulties between study 2009 and study 2000 for the PISA 2000 instrument. Positive values indicate that the item was more difficult in study 2009. Error bars represent standard errors.

Due to a model identification constraint on the item parameters, no SE are estimated for the last item. Item labels are listed in the Appendix.

Science

As only five of the seven PISA 2006 science clusters were used in study 2009, only 44 science link items (out of the full set of 53 items) could be analyzed. Since these five clusters from PISA 2006 were positioned at the fourth cluster position in study 2009, only this cluster position was used for the data from science 2009 as well. The group parameter included in Model 1 revealed that the science link items were easier for participants tested in 2009 with science 2009 than for participants tested in 2009 with science 2006 (-0.27 logits, SE = 0.04). A comparison of the model fit for Model 1 and Model 2 yielded a better fit for Model 2 according to the χ^2-test and the AIC (see Table 4a). Thus, as for reading, the science items also showed an interaction between instruments (2006 vs. 2009) and items, indicating that, differences in item difficulty between the two instruments need to be taken into account. As can be seen in Figure 2, some items differ substantially between science 2006 and science 2009, especially S413Q05 (-3.36 logits, SE = 0.20, 95% CI[-3.75, -2.96]) and S256Q01 (-1.86 logits, SE = 0.39, 95% CI [-2.61, -1.10]) which are both easier in science 2009. When these two items were removed and Model 1 was recomputed, the difference between the two subsamples was reduced to -0.20 logits (SE = 0.04), again favoring students tested with science 2009 in study 2009, though slightly smaller in size compared to the full item set.

The link error for the 42 science link items (without S413Q05 and S256Q01) was 7.55 points on the PISA science scale (see Table 5). This link error is not comparable to the one reported in the PISA 2009 Technical Report (2.57 points; OECD, 2012) which was based on the full set of 53 link items at all four cluster positions. The correlation between the difference in the index for item cluster position and the difference in item difficulty was $r =$ -.03 ($p =$ .84) for the 42 science link items remaining after removal of the two items with the largest differences in item difficulty. The mean difference in item cluster position between science 2009 and science 2006 (for the fourth cluster) was negligible at -0.01 ($SD = 0.18$). Thus, for

the 42 science link items and the fourth cluster position, position effects do not appear to play a role for the differences in item difficulty.

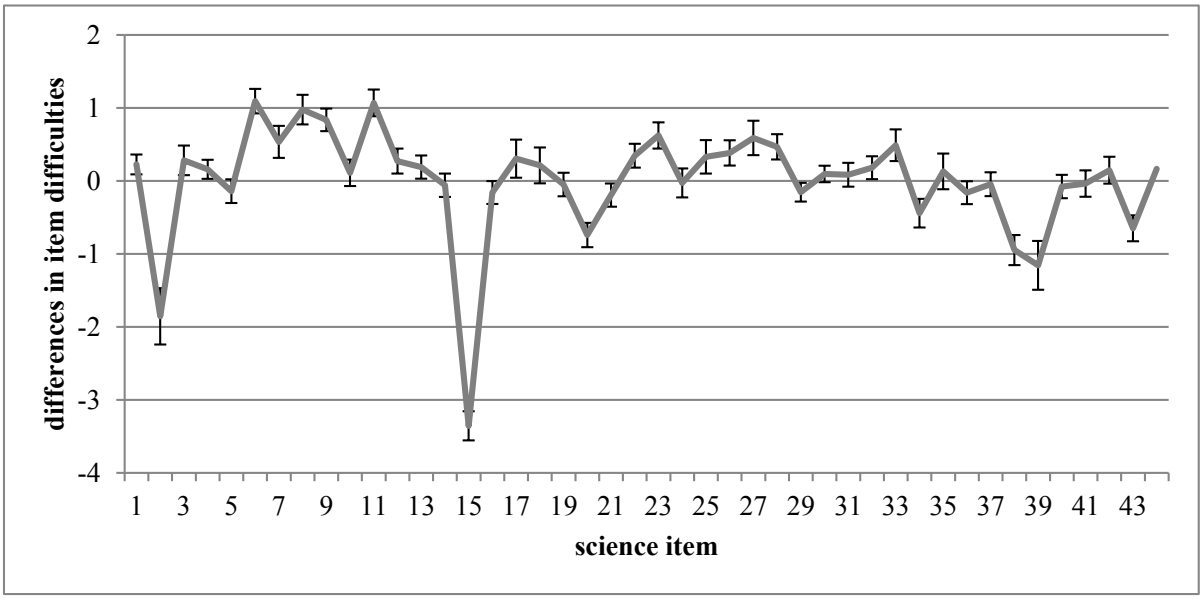

Figure 2. Differences in item difficulties for science between the PISA 2009 instrument and the PISA 2006 instrument. Positive values indicate that the item was more difficult in the PISA 2009 instrument. Error bars represent standard errors.

Due to a model identification constraint on the item parameters, no SE are estimated for the last item. Item labels are listed in the Appendix.

Discussion

In this paper, the measurement invariance – as an important pre-requisite of trend analyses – of PISA reading and science link items was analyzed for items from PISA 2000 and PISA 2009 for reading and from PISA 2006 and PISA 2009 for science. Furthermore, we analyzed whether position effects accounted for differences in item difficulties across instruments and assessments.

Our analyses showed that some of the reading and science link items changed in their difficulty between 2000 and 2009. One possible reason are variations in item wording between the assessments. Regarding the German test booklets applied here, five reading link items (R055Q03, R067Q04, R104Q02, R220Q04, and R227Q02) were phrased slightly differently in PISA 2009 compared to PISA 2000. For R055Q03 this explanation appears especially plausible, since the wording in the German booklets was simplified which may have led to the item being easier in the PISA 2009 instrument compared to the PISA 2000 instrument. As an aside, R055Q03 was deleted on the national level in the German-speaking countries for PISA 2000 and PISA 2003 but has been retained since PISA 2006 (presumably after the wording was changed).

The comparison of model fit for the RCML models making different equality assumptions confirms that differences in item difficulties exist. Thus, the PISA 2000 and PISA 2009 instruments are not measurement invariant regarding the 37 common reading items as well as the 26 reading link items. Furthermore, the instrument from PISA 2000 also was not measurement invariant between two studies (study 2000 and study 2009) regarding the reading items. For 44 of the science link items, measurement invariance was also shown to be violated between the instruments from PISA 2006 and PISA 2009.

The link errors computed with our data were larger compared to the ones reported in the Technical Report for PISA 2009 (OECD, 2012) for the reading link 00/09 and the science link 06/09. However, when items with large differences in item difficulty between the instruments were removed, the link error was reduced by approximately 0 to 3 points on the PISA scale (mean reduction 20.83%). Thus, the few items that changed their difficulties between assessments appear to have had a strong influence on the size of the link error. Furthermore, using the reduced item sets, link errors were larger for the 24 reading link items compared to the 34 common reading items for the link 00/09 in study 2009. For science, the link error

computed from items on cluster position four was much larger than the one computed by the OECD for all cluster positions. For the international sample the link error was smallest at cluster position 1 and largest at cluster position 3. It is conceivable that the link error was increased by fatigue effects for cluster positions 2 and 3. Differences between the link errors from the international samples drawn in this study and the one reported by the OECD are probably due to the different data used: the OECD link error is based on data from all four cluster positions while our link errors are based on data from only the first three cluster positions for reading. Since the link error is computed using the differences in item difficulty between assessments, it can be assumed that differing results on the differences in item difficulty between the OECD sample and our sample contributed to differing link errors. The size of the link error also appears to be influenced by sample size since the link error was slightly larger for an international sample of about 500 students per OECD country compared to an international sample of about 2000 students per OECD country. Large link errors can impair the measurement invariance of PISA instruments and in consequence limit the conclusions that can be drawn from trend analyses. It follows that eliminating factors that lead to large link errors is important. Our results confirm the previous finding by Wu (2010) and Mazzeo and von Davier (2009) that rather more than fewer items should be used to establish the link.

Differences in item position between instruments are a possible explanation for differences in item difficulty. For example, in the PISA 2009 instrument, the reading link items were on average positioned earlier compared to the instrument used in PISA 2000. Thus, these items may have been easier for participants in PISA 2009 due to position effects. Position effects may have played a role for some items. For example, the difference in cluster position and the difference in item difficulties between the instruments from PISA 2000 and PISA 2009 for the reading items showed a small to medium correlation, indicating that on average, reading items were positioned earlier and were easier in reading 2009 compared to reading 2000. It follows

that the recommendation expressed by Mazzeo and von Davier (2009) as well as Wu (2010) of changing as little as possible and assuming that all changes have an effect can only be emphasized as even minor differences between assessments can limit possibilities for trend analyses.

A further factor that influences changes in item difficulty and which in turn enlarges the link error is booklet effects. Booklet effects refer to the position of items in test booklets. According to Wu (2010), link items should be placed at the same position since difficulty changes resulting from position effects may increase the link error. Booklet effects affected item parameter estimates in PISA 2000 (Adams & Carstensen, 2002). Since the test design has been balanced from PISA 2003 on, item parameter estimates in PISA 2006 and PISA 2009 should not be affected by booklet effects. Lastly, carry-over effects may also contribute to difference in item difficulties. This is especially relevant for the comparison between science 2006 and science 2009 as well as reading 2000 and reading 2009 both assessed in study 2009 since here differing items preceded the link and common items we analyzed, possibly contributing to differences in item difficulties, while for the comparison between study 2000 and study 2009 regarding reading the composition of the clusters was identical.

Limitations

The results reported here are based on samples from a single country, namely Germany. Results on the international level or in other countries participating in PISA may differ. The samples used in this study both consisted of high school students. Thus, the results are not generally valid for other school types. Furthermore, while the sample assessed in 2000 was part of the official German PISA 2000 sample, the sample assessed in 2009 formed part of a study conducted by the German PISA consortium in addition to PISA 2009. However, since

this study was conducted in adherence to the PISA procedure (e.g., concerning standardization), it is assumed that the data collection and analyses for this sample do not differ systematically from those of the PISA sample.

Conclusion

The interaction between items and study or instrument, respectively, indicates that measurement invariance between the PISA instruments for 2000 (2006) and 2009 for reading (science) is not given. For some items, differences in item difficulty are substantial. These may partly be attributed to position effects, though other factors play a role as well. For the reading items, the link 2000/2009 works quite well with all common items and shows a smaller link error compared to the link error computed with only the link items. Items with large differences in item difficulty between assessments appear to increase the link error and thus should be removed from linking.

References

Adams, R. J., & Carstensen, C. H. (2002). Scaling outcomes. In OECD, *PISA 2000 Technical Report* (pp. 149–162). Paris, France: OECD.

Adams, R. J., & Wilson, M. R. (1996). Formulating the Rasch model as a mixed coefficients multinomial logit. In G. Engelhard & M. R. Wilson (Eds.), *Objective measurement: Theory into practice. Vol III* (pp. 143–166). Norwood, NJ: Ablex.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & B. F. Csaki (Eds.), *Second international symposium on information theory* (pp. 267–281). Budapest: Academiai Kiado.

Andrich, D. (1978). Application of a psychometric rating model to ordered categories which are scored with successive integers. *Applied Psychological Measurement*, *2*(4), 581–594. doi:10.1177/014662167800200413

Baumert, J., Artelt, C., Klieme, E., Neubrand, M., Prenzel, M., Schiefele, U., … Weiß, M. (Eds.). (2002). *Pisa 2000 - die Länder der Bundesrepublik Deutschland im Vergleich: PISA-E* [PISA 2000 - comparison of the German states: PISA-E]. Opladen, Germany: Leske + Budrich.

Bischof, L., Hochweber, J., Hartig, J., & Klieme, E. (in press). Schulentwicklung im Verlauf eines Jahrzehnts – Erste Ergebnisse des PISA Schulpanels [School development in a decade – First results from the PISA school panel]. *Zeitschrift für Pädagogik*.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 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*(3), 345–370. doi:10.1007/BF02294361

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, *37*, 359–374. doi:10.1016/0001-6918(73)90003-6

Gebhardt, E., & Adams, R. J. (2007). The influence of equating methodology on reported trends in PISA. *Journal of Applied Measurement*, *8*(3), 305–322.

Kolen, M. J., & Brennan, R. L. (2004). *Test Equating, scaling, and linking: Methods and practices* (2nd ed.). New York, NY: Springer.

Linacre, J. M. (1994). *Many-facet Rasch measurement*. Chicago, IL: MESA Press.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174. doi:10.1007/BF02296272

Mazzeo, J. & Davier, M. von. (2009). *Review of the Programme for International Student Assessment (PISA) test design: Recommendations for fostering stability in assessment results*. Retrieved August, 2010, from http://edsurveys.rti.org/PISA.

Monseur, C., & Berezner, A. (2007). The computation of equating errors in international surveys in education. *Journal of Applied Measurement*, *8*(3), 323–335.

OECD. (2002). *PISA 2000 Technical report*. Paris, France: OECD Publications.

OECD. (2009a). *PISA 2009 Assessment framework – Key competencies in reading, mathematics, and science*. Paris, France: OECD Publications.

OECD. (2009b). *PISA 2006 Technical report*. Paris, France: OECD Publications.

OECD. (2010). *PISA 2009. Learning trends: Changes in student performance since 2000*. Paris, France: OECD Publications.

OECD. (2012). *PISA 2009 Technical Report*. Retrieved from http://dx.doi.org/10.1787/9789264167872-en

OECD, & Statistics Canada. (2000). *Literacy in the Information Age: Final report of the International Adult Literacy Survey*. Paris, France: OECD Publications.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danmarks Paedagogiske Institut.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464. doi:10.1214/aos/1176344136

Wu, M. L. (2010). Measurement, sampling, and equating errors in large-scale assessments. *Educational Measurement: Issues and Practice*, *4*(29), 15–27.

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. (2007). ConQuest (Version 2.0) [Computer software]. Camberwell, Australia: Australian Council for Educational Research.

Zieky, M. (1993). Practical questions in the use of DIF statistics in item development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337–364). Hillsdale, NJ: Lawrence Erlbaum.

Appendix 1

Reading and Science Items with Item Parameters (SE) in each Subsample and Mean Position

			Item parameter (SE)			Item parameter (SE)	
Domain	**Item nr.**	**Item label**	**S00 – I00 *N* = 1487**	**S09 – I00 *N* = 554**	**mean position I00**	**S09 – I09 *N* = 1394**	**mean position I09**
Reading	1	R055Q01*	-3.07 (.21)	-2.56 (.20)	2.13	-2.52 (.17)	2.53
	2	R055Q02*	-1.59 (.12)	-1.66 (.15)	2.20	-0.96 (.11)	2.60
	3	R055Q03*	0.48 (.10)	0.16 (.12)	2.27	-1.54 (.13)	2.67
	4	R055Q05*	-2.96 (.20)	-3.25 (.27)	2.33	-2.70 (.19)	2.73
	5	R067Q01*	-4.12 (.34)	-4.06 (.38)	2.60	-4.51 (.42)	2.21
	6	R067Q04*	-0.86 (.08)	-0.60 (.08)	2.67	-0.46 (.07)	2.29
	7	R067Q05*	-1.37 (.10)	-1.19 (.09)	2.73	-1.29 (.09)	2.36
	8	R083Q01	-2.27 (.15)	-1.63 (.18)	2.74	-1.48 (.12)	2.00
	9	R083Q02	-2.73 (.18)	-3.12 (.30)	2.79	-2.71 (.18)	2.06
	10	R083Q03	-2.76 (.18)	-3.03 (.29)	2.85	-2.47 (.17)	2.13
	11	R083Q04	-1.45 (.12)	-1.17 (.16)	2.91	-1.27 (.12)	2.19
	12	R101Q01	-1.46 (.12)	-1.43 (.17)	2.41	-1.18 (.12)	2.69
	13	R101Q02	-3.10 (.21)	-3.32 (.33)	2.47	-2.39 (.16)	2.75
	14	R101Q03	-2.12 (.15)	-1.88 (.19)	2.53	-1.70 (.13)	2.81
	15	R101Q04	-2.20 (.15)	-2.03 (.20)	2.59	-2.27 (.16)	2.88
	16	R101Q05	-0.74 (.11)	-0.29 (.14)	2.65	-0.30 (.10)	2.94
	17	R102Q04A*	-1.40 (.12)	-0.85 (.12)	2.50	-0.52 (.11)	2.43
	18	R102Q05*	-0.42 (.10)	-0.23 (.12)	2.56	-0.46 (.10)	2.50
	19	R102Q07*	-3.46 (.25)	-3.03 (.24)	2.67	-2.93 (.21)	2.57
	20	R104Q01*	-4.26 (.36)	-3.21 (.31)	3.26	-2.73 (.19)	2.80
	21	R104Q02*	0.55 (.10)	0.91 (.16)	3.32	1.50 (.13)	2.87
	22	R104Q05*	0.63 (.11)	1.39 (.24)	3.44	1.69 (.20)	2.93
	23	R111Q01*	-2.08 (.14)	-2.20 (.17)	2.72	-2.26 (.16)	2.27
	24	R111Q02B*	-0.57 (.08)	0.02 (.09)	2.78	0.09 (.08)	2.33
	25	R111Q06B*	-0.43 (.06)	0.04 (.06)	2.94	-0.14 (.06)	2.47
	26	R219Q02*	-2.39 (.16)	-2.55 (.24)	2.11	-2.43 (.17)	2.14
	27	R220Q01*	-1.04 (.11)	-0.21 (.21)	3.69	-0.60 (.11)	2.64
	28	R220Q02B*	-2.14 (.15)	-1.19 (.24)	3.75	-1.69 (.13)	2.71
	29	R220Q04*	-1.87 (.13)	-1.42 (.25)	3.81	-1.58 (.13)	2.79
	30	R220Q05*	-3.73 (.27)	-2.29 (.33)	3.88	-3.45 (.26)	2.86
	31	R220Q06*	-2.21 (.15)	-1.49 (.26)	3.94	-2.21 (.16)	2.93
	32	R227Q01*	-1.00 (.11)	-0.92 (.12)	2.00	-0.92 (.11)	2.00
	33	R227Q02T*	-1.85 (.15)	-1.79 (.17)	2.06	-1.92 (.16)	2.07
	34	R227Q03*	-2.01 (.14)	-1.34 (.14)	2.11	-1.38 (.12)	2.13
	35	R227Q06*	-2.54 (.17)	-2.36 (.19)	2.22	-2.72 (.19)	2.20
	36	R245Q01	-3.01 (.20)	-2.23 (.21)	1.50	-1.95 (.14)	2.56
	37	R245Q02	-3.10 (.21)	-3.42 (.35)	1.58	-2.89 (.20)	2.63

(continued)

			Item parameter (SE)			Item parameter (SE)	
Do-main	**Item nr.**	**Item label**	**S00 – I00 *N* = 1487**	**S09 – I06 *N* = 554**	**mean posi-tion I06**	**S09 – I09 *N* = 1394**	**mean posi-tion I09**
Science	1	S131Q02T	NA	-1.04 (.30)	4.05	-1.10 (.11)	4.18
	2	S256Q01	NA	-2.18 (.36)	4.00	-4.30 (.72)	4.22
	3	S269Q01	NA	-1.94 (.34)	4.18	-1.90 (.29)	4.00
	4	S269Q03T	NA	-0.61 (.29)	4.23	-0.74 (.10)	4.06
	5	S269Q04T	NA	0.66 (.29)	4.27	0.18 (.22)	4.11
	6	S326Q01	NA	-1.45 (.31)	4.10	-0.65 (.22)	4.00
	7	S326Q02	NA	-2.41 (.37)	4.15	-2.14 (.30)	4.06
	8	S326Q03	NA	-2.32 (.36)	4.20	-1.61 (.26)	4.11
	9	S326Q04T	NA	0.09 (.28)	4.25	0.59 (.22)	4.17
	10	S408Q01	NA	-1.28 (.31)	4.30	-1.45 (.26)	4.17
	11	S408Q03	NA	1.03 (.29)	4.35	1.73 (.28)	4.22
	12	S408Q04T	NA	-1.07 (.30)	4.40	-1.09 (.24)	4.28
	13	S408Q05	NA	-0.21 (.28)	4.45	-0.33 (.22)	4.33
	14	S413Q04T	NA	-0.45 (.29)	4.84	-0.82 (.22)	4.50
	15	S413Q05	NA	2.10 (.35)	4.89	-1.59 (.26)	4.56
	16	S413Q06	NA	-0.09 (.29)	4.79	-0.56 (.22)	4.44
	17	S415Q02	NA	-2.74 (.40)	4.90	-2.69 (.38)	4.88
	18	S415Q07T	NA	-2.52 (.38)	4.85	-2.56 (.37)	4.82
	19	S415Q08T	NA	-0.34 (.28)	4.95	-0.69 (.23)	4.94
	20	S425Q02	NA	0.02 (.28)	4.48	-1.04 (.24)	4.89
	21	S425Q03	NA	-0.20 (.29)	4.38	-0.71 (.22)	4.78
	22	S425Q04	NA	-0.61 (.29)	4.52	-0.58 (.23)	4.94
	23	S425Q05	NA	-1.57 (.32)	4.43	-1.23 (.24)	4.83
	24	S428Q01	NA	-1.66 (.33)	4.32	-1.95 (.29)	4.29
	25	S428Q03	NA	-2.40 (.38)	4.37	-2.31 (.32)	4.35
	26	S428Q05	NA	-1.16 (.31)	4.42	-1.07 (.24)	4.41
	27	S438Q01T	NA	-2.61 (.40)	4.53	-2.27 (.32)	4.65
	28	S438Q02	NA	-1.16 (.31)	4.58	-0.98 (.24)	4.71
	29	S438Q03T	NA	0.01 (.29)	4.63	-0.46 (.10)	4.76
	30	S465Q01	NA	-0.39 (.24)	4.16	-0.62 (.15)	4.00
	31	S465Q02	NA	-0.64 (.29)	4.21	-0.86 (.23)	4.06
	32	S465Q04	NA	0.14 (.29)	4.26	-0.01 (.21)	4.12
	33	S466Q01T	NA	-2.20 (.36)	4.79	-1.96 (.31)	4.83
	34	S466Q05	NA	-1.16 (.31)	4.89	-1.87 (.30)	4.94
	35	S466Q07T	NA	-2.40 (.38)	4.84	-2.50 (.37)	4.89
	36	S478Q01	NA	0.23 (.29)	4.37	-0.26 (.21)	4.28
	37	S478Q02T	NA	-0.64 (.29)	4.42	-0.99 (.23)	4.33
	38	S478Q03T	NA	-1.21 (.31)	4.47	-2.43 (.33)	4.39
	39	S514Q02	NA	-2.44 (.38)	4.62	-3.83 (.60)	4.47
	40	S514Q03	NA	-0.20 (.29)	4.67	-0.59 (.22)	4.53

(continued)

			Item parameter (SE)			Item parameter (SE)	
Do-mai n	**Item nr.**	**Item label**	**S00 – I00 *N* = 1487**	**S09 – I06 *N* = 554**	**mean posi-tion I06**	**S09 – I09 *N* = 1394**	**mean posi-tion I09**
Science	41	S514Q04	NA	-1.22 (.31)	4.71	-1.53 (.26)	4.59
	42	S527Q01T	NA	1.54 (.31)	4.55	1.31 (.26)	4.67
	43	S527Q03T	NA	-0.48 (.29)	4.59	-1.42 (.27)	4.72
	44	S527Q04T	NA	-0.57 (.29)	4.64	-0.70 (.24)	4.78

Note. S00 = study 2000. S09 = study 2009. I00 = PISA 2000 instrument. I06 = PISA 2006 instrument. I09 = PISA 2009 instrument.
* reading link items

2.4. Appendix D: Manuscript Wetzel & Carstensen (2013b)

Multidimensional modeling of response styles

Abstract

Response styles can influence item responses in addition to a respondent's latent trait level. Thus, comparisons between individuals based on sum scores may be rendered invalid by response style effects. This paper presents a multidimensional approach to modeling traits and response styles simultaneously in the framework of the multidimensional partial credit model. Models incorporating different response styles (extreme response style, acquiescence, disacquiescence, and midpoint response style) as well as personality traits were compared regarding model fit. Furthermore, relationships between traits and response styles and the correction of trait estimates for response style effects were investigated.

All multidimensional models showed a better fit than the unidimensional models, indicating that response styles influenced item responses. Extreme response style explained more variance in item responses incremental to trait variance than the other response styles. Latent correlations revealed that extreme response style and midpoint response style are mainly trait-independent whereas acquiescence and disacquiescence are strongly related to several traits. Different methods of correcting trait estimates for response style effects are illustrated and discussed.

Keywords: response styles, multidimensional partial credit model, multidimensional modeling, latent correlations

Introduction

Response styles characterize individual differences in response scale use that are independent of item content. Thus, when response styles occur, the choice of a response category is not only influenced by the respondent's trait level but also by his or her tendency to prefer or avoid certain response categories. Common response styles are extreme response style (ERS), the tendency to prefer extreme categories, acquiescence response style (ARS), the tendency to agree, disacquiescence response style (DARS), the tendency to disagree, and midpoint response style (MRS), the tendency to prefer the middle category of a response scale. For a comprehensive review of these response styles see Baumgartner and Steenkamp (2001).

The use of sum scores to draw conclusions concerning a respondent's latent trait level and to draw comparisons between respondents with different sum scores is based on the assumption that the respondents' true latent trait levels are the only influence on item responses. However, when response styles occur, this assumption is violated and in consequence a distortion of sum scores takes place (Austin, Deary, & Egan, 2006; Baumgartner & Steenkamp, 2001). For instance, a person who displays ERS might receive a more extreme score than a person without a preference for extreme categories despite both of them having the same latent trait level.

This paper proposes a multidimensional approach to modeling response styles in which both traits and response styles can be incorporated into the same model. Using this approach, it becomes possible to obtain trait estimates that are corrected for responses styles. In the following, we contrast two common approaches to modeling response styles which are based on a categorical view and a dimensional view of response styles. Then, we present the multidimensional partial credit model (e.g., Kelderman, 1996), which is the underlying model used in our analyses. We apply the multidimensional approach to the standardization sample of the German NEO-PI-R (Ostendorf & Angleitner, 2004) to first investigate whether response styles

explain variance in item responses incremental to the trait variance and if this is the case, which response style explains how much variance. Second, we investigate relationships between response styles and traits and between different response styles. Third, we discuss the correction of trait estimates for response styles in multidimensional models. We end with a discussion of our results and their implications.

Modeling Response Styles: The Categorical and the Dimensional Approach

Two differing views exist regarding the nature of response styles: some researchers see response styles as categorical variables where people either have a response style or they do not (*categorical approach;* e.g., Austin et al., 2006) while other researchers see response styles as continuous variables where the response style is a dimension on which persons differ in the degree to which they show a certain response style (*dimensional approach;* e.g., Greenleaf, 1992a). In the following, only studies that model response styles in the framework of Item Response Theory will be considered. For approaches to investigating response styles in the framework of Classical Test Theory see for example Baumgartner and Steenkamp (2001) or Greenleaf (1992b) or, for an alternative modeling approach based on a Bayesian hierarchical model, see Rossi, Gilula, and Allenby (2001).

Analyses following the categorical approach usually apply mixed Rasch models (Rost, 1990, 1991) to differentiate latent classes that systematically differ in their response scale use. Mixed Rasch models allow the investigation of both qualitative differences between subgroups of participants (i.e., latent classes) as well as quantitative differences between the participants in one subgroup (i.e., in each latent class, the Rasch model holds). The latent classes resulting from a mixed Rasch analysis are then interpreted as different response styles using the distribution of the threshold parameters. For example, Rost, Carstensen, and von Davier (1999)

examined the German NEO-FFI (Borkenau & Ostendorf, 1993) and found that for neuroticism, openness to experience, agreeableness, and conscientiousness two-class solutions were adequate to describe the data. These two classes could be inter-preted as a class of extreme responders (ERS) and a class of participants who used the response scale evenly or displayed a preference for moderate response categories (non-extreme response style, NERS). The same result was found by Austin et al. (2006) in the English NEO-FFI (Costa & McCrae, 1992) concerning neuroticism, extraversion, agreeableness, and conscientiousness and by Eid and Rauber (2000) in an analysis of a leadership performance scale. Maij-de Meij, Kelderman, and van der Flier (2008) found three latent classes in the extraversion and neuroticism scales of the Amsterdam Biographical Questionnaire which differed regarding social desirability, ethnic background, and the use of the "?" category. Differential use of the "?" category was also shown by Hernández, Drasgow, and González-Romá (2004) to be the defining characteristic of the two latent classes derived for most of the non-cognitive scales in the 16PF (Cattell, Cattell, & Cattell, 1993). Thus, in the categorical approach the underlying assumption made is that participants can be divided into two groups, those who use a certain response style and those who do not, and that participants in these two groups differ qualitatively.

In the dimensional approach response styles are modeled as continuous variables. For example, Bolt and Johnson (2009) suggested modeling ERS in addition to the trait of interest in a multidimensional extension of Bock's (1972) nominal response model. They applied this model to a measure of tobacco dependence and found that incorporating the trait (level of tobacco dependence) and a response style dimension (ERS) into the same model allowed examining the impact of response styles on trait scores as well as identifying whether the source of differential item functioning was response styles or other factors. Bolt and Newton (2011) extended this approach by simultaneously modeling two traits and one ERS dimension. Using

data from two attitude scales included in the Programme for International Student Assessment's (PISA) 2006 student questionnaire, they showed that including a second trait into the model facilitates the estimation of a person's standing on the ERS dimension since it makes it easier to distinguish whether extreme responses are indicative of a high level on the first trait or a high level of ERS. They argue that it is then possible to estimate the trait level more precisely.

Our study is similar to Bolt and Newton (2011) in that we incorporate response styles and traits in a multidimensional model but it goes beyond Bolt and Newton by investigating more response styles than only ERS and by incorporating more traits as well as more than one response style in multidimensional models. Furthermore, we examine the relationship between traits and response styles on the one hand and different response styles on the other hand using latent correlations. Lastly, we also elucidate the issue of obtaining trait estimates that are corrected for response style effects in more detail. Our study draws upon results from previous analyses of the same sample presented in Wetzel, Böhnke, Carstensen, Ziegler, and Ostendorf (in press) and Wetzel, Carstensen, and Böhnke (2013). In the first study Wetzel et al. (in press) conducted analyses of the German NEO-PI-R's standardization sample according to the categorical approach and found that on many of the NEO-PI-R facets two latent classes could be differentiated. These were interpreted as extreme response style and non-extreme response style. In the second study, Wetzel et al. (2013) entered manifest class memberships to ERS or NERS on each scale into a latent class analysis to investigate the consistency of response styles across the scales in an instrument. For the majority of the participants (between 65 and 80%) membership to either the ERS or the NERS class was consistent across scales in two instruments (PISA 2006 attitude scales and NEO-PI-R facets). It follows that respondents who preferred or avoided extreme categories on one scale also tended to do so on the other scales. In addition, the latent classes showed mainly quantitative differences in their allocation to ERS

or NERS, while qualitative differences were not apparent. Thus, modeling response styles as continuous variables appears to be a promising alternative to the categorical approach that will be investigated in this study.

The Multidimensional Partial Credit Model

In this study, the multidimensional partial credit model (MPCM) will be used to model the traits assessed by the NEO-PI-R and the different response styles. The unidimensional Rasch model (Rasch, 1960) and the partial credit model (Masters, 1982) were extended to multidimensional data by the work of several researchers, though the original multidimensional Rasch model dates back to Rasch (1961). For instance, Andersen (1985) and Embretson (1991) developed models for the multidimensional measurement of change in longitudinal data. Stegelmann (1983) and Carstensen (2000) also developed multidimensional models estimated via conditional maximum likelihood while Adams, Wilson, and Wang (1997) and von Davier (2005) derived marginal maximum likelihood estimators for their multidimensional models. Glas and Verhelst (1995) introduced a multidimensional model with both conditional and marginal maximum likelihood estimation. A direct extension of the partial credit model to multidimensional data was presented by Kelderman (1996). In the following, the notation introduced by Kelderman (1996) will be used.

In the MPCM, s trait parameters θ_{jq} ($q = 1,\ldots, s$) exist. w_{qiy} is an indicator variable prespecified by the researcher which reflects the item-dimension relationship. It takes the value 1 if the response to that item represents dimension θ_{jq} and 0 if it does not. The MPCM then models the probability (π_{ijx}) that a person j with trait levels θ_{jq} on the s dimensions will respond in category x ($x = 1, \ldots, r$) of item i as

$$\pi_{ijx} = \frac{\exp\left[\sum_{y=1}^{x}\left(\sum_{q=1}^{s} w_{qiy}\theta_{jq} - \delta_{iy}\right)\right]}{1+\sum_{z=1}^{r_i}\exp\left[\sum_{y=1}^{z}\left(\sum_{q=1}^{s} w_{qiy}\theta_{jq} - \delta_{iy}\right)\right]} \quad (1)$$

where $\sum_{y=1}^{0}(\cdot) \equiv 0$. In Equation 1, δ_{iy} denotes the threshold parameter between two response categories x-1 and x. When $s = 1$ and $w_{qiy} = 1$ in Equation 1, the unidimensional PCM results.

Research Questions

The goal of this study is to investigate three research questions related to the multidimensional modeling of response styles:

1) Do response styles explain additional variance in item responses?

Using model comparisons between unidimensional and multidimensional models we investigate whether response styles can explain variance in item responses in addition to the trait. Furthermore, we compare four different response styles (ERS, ARS, DRS, and MRS) with respect to how much variance in item responses they can explain. Drawing upon previous results (e.g., Austin et al., 2006; Bolt & Johnson, 2009; Rost et al., 1999; Wetzel et al., in press), we expect that response styles explain variance in item responses. In particular, we expect ERS to be the response style that explains the most variance in item responses incremental to the trait.

2) How are response styles and traits as well as different response styles related to each other?

Using latent correlations we investigate relationships between response styles and traits in two-dimensional models and relationships between different response styles in three-dimensional models. Previous studies found that ERS was related to different traits (extraversion and conscientiousness in Austin et al., 2006; intolerance of ambiguity and simplistic thinking in Naemi, Beal, & Payne, 2009). Thus, we expect ERS to be related to several NEO-PI-R facets as well, especially on extraversion and conscientiousness. Analyses on relationships between traits and other response styles as well as between different response styles were exploratory.

3) How does the correction of trait estimates for response styles in multidimensional models change with different response style indicators?

We assume that trait estimates derived from appropriate multidimensional models are corrected for response style effects. Bolt and Newton (2011) found such a corrective effect in three-dimensional models. We propose models with and without the correction of trait estimates and argue that the correction of trait estimates depends on how response styles are measured. This is shown by comparing models based on different response style indicators. A related issue is whether different indicators for response styles assess the same response style and are thus equivalent (i.e., the response style is independent from the trait items used to measure it) or whether there is a scale-specific component to the response style (i.e., the response style differs depending on which trait items are used to measure it). Which is the case carries implications for the correction of trait estimates. Furthermore, we illustrate the correction of trait estimates with a method based on residualized scores.

Method

Sample

The sample comprised the non-clinical standardization sample of the German NEO-PI-R (Ostendorf & Angleitner, 2004). In total, 11,724 persons were part of the sample with 64% women. Participants' age ranged from 16 to 91 years (M = 29.92, SD = 12.08). Cronbach's alpha coefficients for sum scores on the Big Five higher order domains were .93 for neuroticism, .89 for extraversion, .89 for openness to experience, .87 for agreeableness, and .90 for conscientiousness.

Instrument

The NEO-PI-R (Costa and McCrae, 1992) was applied in the German version developed by Ostendorf and Angleitner (2004). The NEO-PI-R assesses the Big Five personality domains (neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness) on the higher-order level as well as on a lower-order facet level. For example, neuroticism can be divided into the six facets anxiety, angry hostility, depression, self-consciousness, impulsiveness, and vulnerability. The NEO-PI-R contains 240 items, eight for each of the 30 facets.

Modeling Traits and Response Styles

In general, variance in responses to questionnaire items using Likert-type scales can be decomposed into trait variance, error variance, and response style variance. In unidimensional models, trait and response style variances are confounded. To achieve a correct estimation of trait variance, variance due to response styles should be taken into account by introducing further latent variables as additional dimensions into the item response model. For interpretational as well as estimating convenience, the trait and response style dimensions of the model should be rather independent of each other. This independence can be achieved by an appropriate coding scheme that codes item responses differently for trait and response style dimensions.

In the literature, several methods of operationalizing response styles and thus separating trait and response style dimensions have been proposed. For example, Meiser and Böckenholt (2011; see also Böckenholt, 2012) posit that the response process consists of multiple steps that can be modeled using different pseudo-items. This framework distinguishes between trait-related processes, during which the respondent decides whether to agree or disagree to the item, and response styles processes, during which the respondent decides whether

to choose an extreme category (i.e., ERS), a moderate category, or the neutral middle category (i.e., MRS). The pseudo-items are dichotomously scored to indicate which decision the respondent made during each sub-process. A sequence of graded response models (Samejima, 1969) is then applied to model the complete response process. With these separate pseudo-items for sub-processes related to the trait and response styles, Meiser and Böckenholt ensure that the dimensions for trait and response styles are largely independent of each other.

We apply a different approach to coding trait dimensions and response style dimensions from the observed item responses. The coding assigns scores to each response category that indicate the trait and response style, respectively (see for example the ordered partition model by Wilson, 1992). The scoring functions applied in this study are depicted in Table 1. For traits, the scores were equivalent to the numerical values of the response categories *strongly disagree* to *strongly agree* (i.e., 0, 1, 2, 3, 4 or for negatively worded items 4, 3, 2, 1, 0). For response styles, the original item responses were scored differently to represent each response style. For example, to model ERS1 we scored the extreme categories with 1 whereas moderate categories were scored with 0. Table 1 shows how item responses were scored to obtain indicators for the other response styles. Thus, our approach achieves a separation of traits and response styles by assigning different scores to indicate traits and response styles. For ERS and MRS, this method of modeling traits and response styles largely achieves independent trait and response style dimensions since the scoring vectors are independent of each other (0, 1, 2, 3, 4 vs. 1, 0, 0, 0, 1 for ERS and 0, 0, 1, 0, 0 for MRS, respectively). However for ARS and DRS the scores for the trait and response style dimensions are more dependent (0, 1, 2, 3, 4 vs. 0, 0, 0, 1, 1 for ARS and 1, 1, 0, 0, 0 for DRS, respectively). This implies that for ARS and DRS, stronger relationships to the trait dimensions can be expected. Note that even with rather independent scorings for traits and response styles the observed data may imply dependencies between both types of latent variables.

Though Meiser and Böckenholt's (2011) procedure of constructing pseudo-items may lead to more strongly independent dimensions, it has the disadvantage that they also dichotomize the trait-related sub-process by scoring both response categories stating agreement with 1 and both response categories stating disagreement with 0. Our method retains the original response scale in the scoring of the trait dimension and thus achieves a more finely graded operationalization of the trait.

Bolt and Newton (2011) applied an operationalization of ERS similar to ours: they coded the ERS dimension using negative weights for extreme responses. For purposes of comparison with Bolt and Newton we also illustrate this option in our analyses, but note that it is formally equivalent to ERS1 (see ERS-1 in Table 1). Additionally, we used two different coding versions for ERS. ERS1 represents the most straightforward way of coding ERS in that extreme responses are weighted with one while non-extreme responses are assigned a score of zero. ERS2 differs from ERS1 and ERS-1 in that the moderate categories also receive a score unequal zero. Thus, ERS2 represents a more fine-grained operationalization of ERS. ERS1 and ERS2 were compared with respect to which would be the most adequate representation of ERS. Due to the use of three scores in ERS2, it should be more strongly related to the trait dimension. Thus, we expect that ERS1 will function best in terms of yielding the largest increment in explained variance.

Another subject of debate in the modeling of response styles is whether to use the same items that are used for the trait dimension (Bolt & Newton, 2011) or whether to use items from different traits (e.g., Baumgartner & Steenkamp, 2001). Both modeling alternatives can be implemented in multidimensional models and are applied in this paper. Figure 1 illustrates these modeling alternatives for ten exemplary items. Items 1 to 6 load on both the trait and the response style. Thus, the trait and response style are modeled simultaneously using the exact same items. In multidimensional models that apply this approach, we expect that the variance

estimated for the trait will be free from response style effects since variance due to the response style is accounted for in the model. This implies that the trait variance reflects the pure trait variance and in consequence, trait estimates derived from these models will be corrected for response style effects. On the other hand, if another item set is added to the model that only measures the response style dimension (as illustrated by items 7 to 10 in Figure 1), the variance due to response styles is no longer measured by the same items as the trait variance. Measuring the response style with additional items may change the response style and in turn may change the correction of trait estimates in the model. When the item sets used to model the trait and the response style show no overlap (i.e. measuring the trait with items 1 to 6 in Figure 1 and the response style with items 7 to 10 only), no correction of trait estimates is implied by the model. In this case, trait variance will be confounded with response style variance. To substantiate this line of reasoning, we will contrast our results on the correction of trait estimates between models that use the same items to model both traits and response styles, models that include other items in addition to the trait items to model response styles, and models that use separate item sets to model traits and response styles. These three types of models will be referred to as 1) *modeling response styles with the same items*, 2) *modeling response styles with additional items*, and 3) *modeling response styles with separate items* in the following.

Table 1

Scoring functions for Trait and Response Style Dimensions

	Response category				
Dimension	**0**	**1**	**2**	**3**	**4**
Trait	0	1	2	3	4
Trait (negatively worded item)	4	3	2	1	0
ERS1	1	0	0	0	1
ERS2	2	1	0	1	2
ERS-1	1	-1	-1	-1	1
ARS	0	0	0	1	1
DRS	1	1	0	0	0
MRS	0	0	1	0	0

Note. ERS = extreme response style, ARS = acquiescence response style, DRS = disacquiescence response style, MRS = midpoint response style.

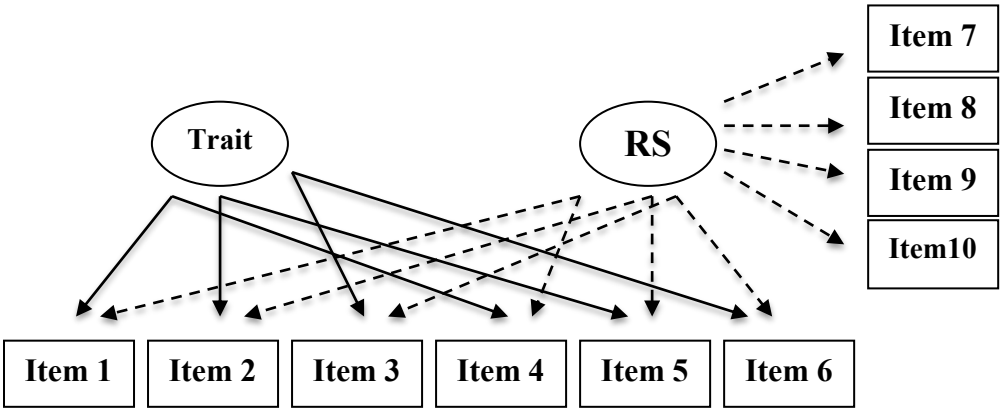

Figure 1. Modeling of trait and response style dimensions.

RS = response style, solid line = scoring for trait dimension (0, 1, 2, 3, 4), dashed line = scoring for response style dimension (e.g., 1, 0, 0, 0, 1 for ERS1).

Trait Estimates in the Multidimensional Partial Credit Model

ConQuest (Wu, Adams, Wilson, & Haldane, 2007), the software used for the estimation of the models in this study, applies marginal maximum likelihood estimation using an EM

algorithm (Bock & Aitkin, 1981) to provide estimates of item parameters and a normally distributed ability density. In the first stage of estimation, latent correlations between the dimensions are estimated and item parameters are estimated to reflect these relationships between dimensions. In the second stage, using the item parameters obtained in the first stage, different trait estimates for the dimensions can be obtained, such as expected a posteriori estimates, plausible values, or weighted likelihood estimates (WLE; Warm, 1989). In this paper, WLEs will be used to illustrate how information on participants' trait levels on all dimensions in the MPCM are taken into account in the estimation of WLEs for one dimension and how this modeling approach can be utilized to obtain trait estimates that are corrected for response style effects.

Analyses

1) Explanation of variance by response styles

First we investigated whether response styles could explain variance in item responses in addition to the trait. To this purpose, different multidimensional models were estimated which included either one trait and one response style or one trait and two response styles. In the two-dimensional models either ERS (in one of the coding variations ERS1 or ERS2), ARS, DRS, or MRS were included as the response style dimension in addition to one of the NEO-PI-R facets as the trait dimension. These four different response styles were used to examine which response style increased the amount of explained variance in item responses the most. In the three-dimensional models different combinations of two response styles were modeled (e.g., ERS and DRS or ARS and MRS) to investigate whether this would improve model fit further.

The unidimensional partial credit model, in which only one trait (a NEO-PI-R facet) was included in each model, was used as the baseline model to which the multidimensional

models were compared. If the unidimensional model fit better than the multidimensional models, we could conclude that response styles do not have a notable influence on item responses. On the other hand, if multidimensional models fit better, we could conclude that response styles can explain variance in item responses in addition to the trait. We hypothesize that the latter will be the case. As ERS is the response style most commonly found in mixed Rasch analyses (e.g., Austin et al., 2006) as well as in multidimensional analyses (e.g., Bolt & Johnson, 2009), we expect ERS to lead to the largest increment in explained variance (i.e., to be the most important response style).

Model fit was assessed using the information criteria Akaike's Information Criterion (AIC; Akaike, 1973), Bayesian Information Criterion (BIC; Schwarz, 1978), and Consistent Akaike's Information Criterion (CAIC; Bozdogan, 1987). The lowest AIC, BIC, and CAIC values indicate the relatively best-fitting model.

2) Relationships between response styles and traits and between different response styles

In the multidimensional partial credit model it is possible to obtain unbiased estimates (i.e., without measurement error) of the true correlations between the latent variables (Adams et al., 1997; Andersen, 1985; Bollen, 1989; Wang, 1999). For the two-dimensional models latent correlations between traits and response styles will be used to investigate the relationships. Latent correlations from the three-dimensional models (one trait, two response styles) will be examined to investigate how different response styles are associated.

3) Correction of trait estimates in the multidimensional partial credit model

If response styles play a role in influencing the choice of a response category, trait estimates from unidimensional models and models that incorporate a response style dimension

should differ. This will be investigated using correlations between weighted likelihood trait estimates from unidimensional PCMs and MPCMs that include one or two traits as well as a response style dimension. Furthermore, MPCMs that include a response style dimension that is based on the same items as the trait dimension should provide trait estimates that are corrected for response style effects. As depicted above, a correction should not take place when separate item sets are used to model traits and response styles. This will be examined by comparing the WLEs from different models exemplarily for some cases and by contrasting the relationship between sum scores on the trait and response style dimensions with the relationship between WLEs on the trait and response style dimensions. We will contrast models with different response style indicators to show in which models a correction of trait estimates occurs. To this purpose, WLEs will be compared between 1) two-dimensional models that use the same items for the trait and response style dimensions (e.g., eight items from a NEO-PI-R facet model both dimensions; modeling response styles with the same items), 2) multidimensional models that include several trait dimensions and one response style dimension (e.g., three traits are modeled by eight items each and the response style dimension is modeled by all 24 items; modeling response styles with additional items), and 3) models that use separate item sets for the two dimensions (e.g., eight items from one facet model the trait and eight items from another facet model the response style; modeling response styles with separate items). To illustrate that whether a correction of trait estimates occurs or not is only a question of the modeling method, a combination of the first and third types of models will also be applied. That is, the same eight items from one facet will be treated as two separate sets of items: one to model the trait and one to model the response style.

Results

1) Explanation of variance by response styles

A direct comparison of our results with Bolt and Newton (2011) would require the operationalization of ERS with negative weights (i.e., ERS-1). However, convergence problems occurred during the estimation of models with ERS-1 for some NEO-PI-R facets which may be attributed to software restrictions. That is, convergence problems can occur in ConQuest when the sum of item response frequencies (i.e., the sufficient statistics) results in a negative value which was the case in our analyses due to the use of negative weights. Thus, ERS-1 was not used in the following model comparisons. Nevertheless, ERS1 and ERS-1 are formally equivalent and should yield the same results.

Exemplarily for anxiety, the comparison between the unidimensional partial credit model (with the trait) and the two-dimensional and three-dimensional MPCMs (including one or two response styles in addition to the trait) is shown in Table 2. All multidimensional models fit better than the unidimensional model according to AIC, BIC, and CAIC. For anxiety the best-fitting model was the three-dimensional PCM with the dimensions anxiety, ERS, and MRS.

Over the 30 NEO-PI-R facets, ERS1 showed the best fit for 21 facets and ERS2 for nine facets when only the unidimensional models (one trait each) and two-dimensional models (one trait and one response style each) are taken into account (see histogram in Figure 2). For the two-dimensional models the largest drop in the AIC, BIC, and CAIC values was observed for models including one of the two operationalizations of ERS, indicating that ERS - especially in its operationalization as ERS1 - was more important in explaining variance in item responses than ARS, DRS, or MRS. These results confirm our hypothesis that ERS is the response style that can explain the most variance in item responses incremental to the trait.

Table 2

Model Fit Comparison for Anxiety

Nr. of dim.	Model	-2 logL	Npar	AIC	BIC	CAIC
1	PCM	238855.92	33	238921.92	239165.11	239198.11
2	ERS1	235502.60	35	235572.60	235830.52	235865.52
2	ERS2	235201.25	35	235271.25	235529.17	235564.17
2	ARS	238646.20	35	238716.20	238974.13	239009.13
2	DRS	238434.22	35	238504.22	238762.15	238797.15
2	MRS	237233.11	35	237303.11	237561.04	237596.04
3	ERS1_ARS	235164.95	38	235240.95	235520.98	235558.98
3	**ERS1_MRS**	**233888.24**	**38**	**233964.24**	**234244.27**	**234282.27**
3	ERS1_DRS	235117.63	38	235193.63	235473.67	235511.67
3	MRS_ARS	237051.182	38	237127.18	237407.22	237445.22
3	MRS_DRS	237051.322	38	237127.32	237407.36	237445.36
3	ARS_DRS	237057.785	38	237133.78	237413.82	237451.82

Note. ERS = extreme response style, ARS = acquiescence response style, DRS = disacquiescence response style, MRS = midpoint response style, dim = dimension, logL = loglikelihood, npar = number of parameters, AIC = Akaike's Information Criterion, BIC = Bayesian Information Criterion, CAIC = Consistent Akaike's Information Criterion. For all models the trait dimension is the anxiety facet. The best-fitting model is depicted in boldface.

Since ERS1 was the coding version of ERS that showed the best fit, three-dimensional models including ERS were only estimated using the coding version ERS1. If the three-dimensional models are additionally taken into account in the model comparison, the model with one trait, an ERS1 dimension, and an MRS dimension fit best for 26 facets, the model with one trait, an ERS1 dimension, and an ARS dimension fit best for three facets, and the model with one trait, an ERS1 dimension, and a DRS dimension fit best for one facet according to all three information criteria. Three-dimensional models without ERS1 (i.e., ARS and MRS, ARS and DRS, MRS and DRS) never showed the best fit. This confirms the conclusion drawn from the two-dimensional models that ERS appears to be the most important response style. Modeling MRS in addition to ERS led to the largest increment in explained variance. In sum, the model comparisons confirmed our expectations that response styles influence item responses in addition to the trait with especially ERS playing an important role.

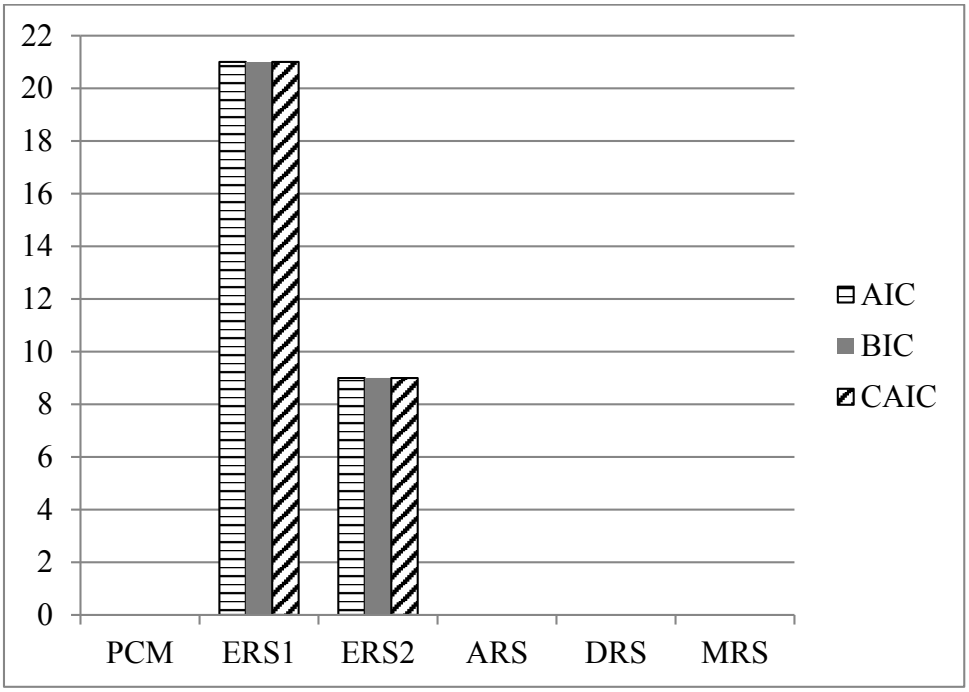

Figure 2. Histogram for the model fit comparison between the unidimensional partial credit model and the two-dimensional models.

ERS = extreme response style (for the two versions see explanation in text), ARS = acquiescence response style, DRS = disacquiescence response style, MRS = midpoint response style AIC = Akaike's Information Criterion, BIC = Bayesian Information Criterion, CAIC = Consistent Akaike's Information Criterion.

2) Relationships between response styles and traits and between different response styles

Latent correlations between the traits assessed by the NEO-PI-R and four different response styles from the two-dimensional MPCMs are depicted in Table 3. Over the 30 NEO-PI-R facets, latent correlations between traits and ERS1 or MRS are mainly negligible, with the exception of the small to moderate (Cohen, 1988) negative correlations between altruism and ERS1 (r = -.205), compliance and ERS1 (r = -.309), modesty and ERS1 (r = -.201), openness to fantasy and MRS (r = -.265), and openness to feelings and MRS (r = -.207). Thus, our hypothesis that ERS would be related to the NEO-PI-R traits was only partly confirmed. For

ARS and DRS a different picture emerges. Here most facets are related to response styles, with some of them showing large correlations (e.g., for achievement striving r = .824 with ARS and r = -.707 with DRS). Correlations for ARS and DRS are in opposite directions with very few exceptions (e.g., openness to feelings is positively associated with both ARS and DRS).

Latent correlations between different response styles derived from three-dimensional models were averaged across the six facets of each Big Five domain. ERS1 and ARS showed small to moderate positive relationships whereas ERS1 and MRS showed small negative relationships (see Table 4). ERS1 and DRS were not associated. ARS and DRS correlated highly and ARS and DRS were also both moderately associated with MRS.

In the NEO-PI-R, between two and five of eight items on each scale are negatively worded. To explore whether the number of negatively worded items had an effect on the correlations between traits and response styles and between different response styles we regressed the absolute values of the correlations on the number of negatively worded items on the respective facet. None of the regressions yielded a significant result, though for correlations between traits and ARS and MRS, the amount of explained variance was not negligible (for ARS R^2=.053, $F(1,28) = 1.589$, p = .219 and for MRS R^2=.052, $F(1,28) = 1.547$, p = .224). The correlation between ARS and MRS increased with rising numbers of negatively worded items with an R^2 of .035 ($F(1,28) = 1.014$, p = .323).

Table 3

Latent Correlations between Traits and Response Styles from Two-dimensional Models

	ERS1	ARS	DRS	MRS
Neuroticism				
N1 Anxiety	.030	-.254	.483	-.134
N2 Angry hostility	.162	-.593	.589	-.067
N3 Depression	.184	-.445	.535	-.086
N4 Self -consciousness	.065	-.083	.274	-.104
N5 Impulsiveness	.063	.212	-.012	-.121
N6 Vulnerability	-.039	-.505	.356	.098
Extraversion				
E1 Warmth	.003	.087	.398	-.072
E2 Gregariousness	.144	.091	-.015	-.027
E3 Assertiveness	.039	-.102	.081	.027
E4 Activity	.098	.113	.096	-.097
E5 Excitement-seeking	.147	-.143	.105	.022
E5 Positive emotions	.037	.856	-.819	-.024
Openness to Experience				
O1 Fantasy	.180	.776	-.523	-.265
O2 Aesthetics	-.067	-.005	.070	.034
O3 Feelings	.107	.206	.470	-.207
O4 Actions	-.038	-.209	.132	.114
O5 Ideas	.152	-.014	.160	-.080
O6 Values	-.156	.334	-.232	.054
Agreeableness				
A1 Trust	-.143	.213	-.272	.015
A2 Straightforwardness	-.030	-.196	.479	-.089
A3 Altruism	-.205	.139	-.171	.125
A4 Compliance	-.309	-.576	.314	.184
A5 Modesty	-.201	-.526	.471	.084
A6 Tender-mindedness	-.095	.311	.019	-.098
Conscientiousness				
C1 Competence	-.097	.637	-.609	-.127
C2 Order	.107	.662	-.275	-.183
C3 Dutifulness	.098	.397	-.109	-.122
C4 Achievement striving	-.081	.824	-.707	-.042
C5 Self-discipline	-.070	.261	-.244	.020
C6 Deliberation	.019	.422	-.133	-.128

Note. ERS1 = extreme response style, ARS = acquiescence response style, DRS = dis-acquiescence response style, MRS = midpoint response style.

Table 4

Latent Correlations between Different Response Styles from Three-dimensional Models

		Response style		
Domain		**ERS1**	**ARS**	**DRS**
Neuroticism	**ARS**	.201		
	DRS	.070	.764	
	MRS	-.188	.413	.296
Extraversion	**ARS**	.305		
	DRS	-.021	.819	
	MRS	-.165	.421	.192
Openness	**ARS**	.126		
	DRS	.035	.672	
	MRS	-.136	.449	.385
Agreeableness	**ARS**	.232		
	DRS	-.023	.732	
	MRS	-.167	.410	.343
Conscientiousness	**ARS**	.128		
	DRS	.179	.781	
	MRS	-.221	.485	.194

Note. ERS = extreme response style, ARS = acquiescence response style, DRS = disacquiescence response style, MRS = midpoint response style. Latent correlations were averaged across the six facets of each Big Five domain.

3) Correction of trait estimates in the multidimensional partial credit model

Results concerning the estimation of WLEs on trait and response style dimensions will only be reported for ERS1 since this emerged from the model fit comparisons as the most important response style. Especially for the two-dimensional models (one trait and ERS1), ConQuest encountered convergence problems in the estimation of WLEs for a large amount of respondents (persons with extreme responses on three or more of the eight items on one facet). For the three-dimensional models (two traits and ERS1) this was the case less often. Cases with convergence problems in WLE estimation were excluded from the computation of correlations between WLEs. No convergence problems occurred during the estimation of models that used separate item sets for the trait and response style dimensions.

Model comparisons reported above showed that multidimensional models including response styles fit better than unidimensional models. The occurrence of a change in the trait estimates between unidimensional and multidimensional models can be derived from the reduced trait variance in multidimensional models compared to unidimensional models. On average across the 30 facets, the trait variance was reduced by 10.17% in the two-dimensional models with one trait and one ERS dimension compared to the unidimensional models with only one trait dimension.

To illustrate the change in trait estimates when ERS is added to the model, trait WLEs derived from the unidimensional model and the two-dimensional model that additionally incorporates ERS (modeling ERS with the same items) were correlated. Across the 30 facets, correlations were high but not perfect (average $r = .85$). Figure 3 shows this exemplarily for self-consciousness. While WLEs for self-consciousness from the unidimensional PCM and the two-dimensional PCM showed a strong association for many cases, deviations in both directions were also numerous. Hence, for some people the WLE in the two-dimensional model which includes a response style dimension is adjusted upwards or downwards compared to their WLE in the unidimensional PCM. This illustrates the "correction" of trait variance from response style variance, when the same items are used to model both the trait and the response style.

Table 5 exemplarily shows sum scores, response patterns, and WLEs for two cases on anxiety and angry hostility. The two cases have the same sum score (20) on anxiety but differ strongly in their tendency to endorse extreme categories, case number 8 used extreme responses very rarely while case number 9700 has a large amount of extreme responses (see sum score ERS1 in Table 5). This allows comparing the trait estimates derived from different models with respect to whether response style effects are corrected for or not. In the upmost part

of the table, WLEs obtained from separate unidimensional PCMs for anxiety and angry hostility are shown. Here, persons with the same sum score also receive the same WLE (0.57 on anxiety for a sum score of 20). This is not the case when WLEs are estimated in a multidimensional model which includes anxiety and angry hostility (or more neuroticism facets) as well as an ERS1 dimension based on the combined items from all traits (modeling ERS with additional items). Here, respondents' WLEs on anxiety are adjusted depending on their levels on the other traits and their ERS levels, resulting in different WLEs for these respondents despite them having the same sum score. Furthermore, the adjustment of WLEs in multidimensional models that use the same items to model both trait and response style dimensions leads to trait estimates that are corrected for response styles. For example, case 7600 gave four extreme responses on anxiety (out of 8 items). Accordingly, this person's anxiety WLE is lower in the multidimensional models compared to the unidimensional model since information on the person's ERS tendency (high ERS WLE) is also taken into account during estimation. Figure 4 illustrates the corrective effect taking place in WLEs across the complete sample by contrasting sum scores with WLEs. Figure 4a shows the relationship between trait sum scores and ERS sum scores for angry hostility. ERS sum scores are distributed almost evenly across all levels of the angry hostility sum scores, though of course extreme trait sum scores are only possible with extreme ERS sum scores. In contrast, for angry hostility WLEs and ERS WLEs estimated in the two-dimensional model, a negative relationship can be observed (Figure 4b). Thus, respondents with high levels of ERS tend to receive lower trait estimates, indicating that trait estimates are adapted for the respondents' ERS. However, with more traits being included in these multidimensional models and correspondingly more items being used to build the ERS dimension, the corrective effect on the trait estimates decreases. For instance, for case number 8, the more items are used to model ERS, the less the proportion of extreme responses becomes (sum score ERS1 in Table 5) which in turn leads to rising anxiety WLEs. Moreover, WLEs

for the ERS dimension fluctuate strongly depending on the number of items used to model the ERS dimension.

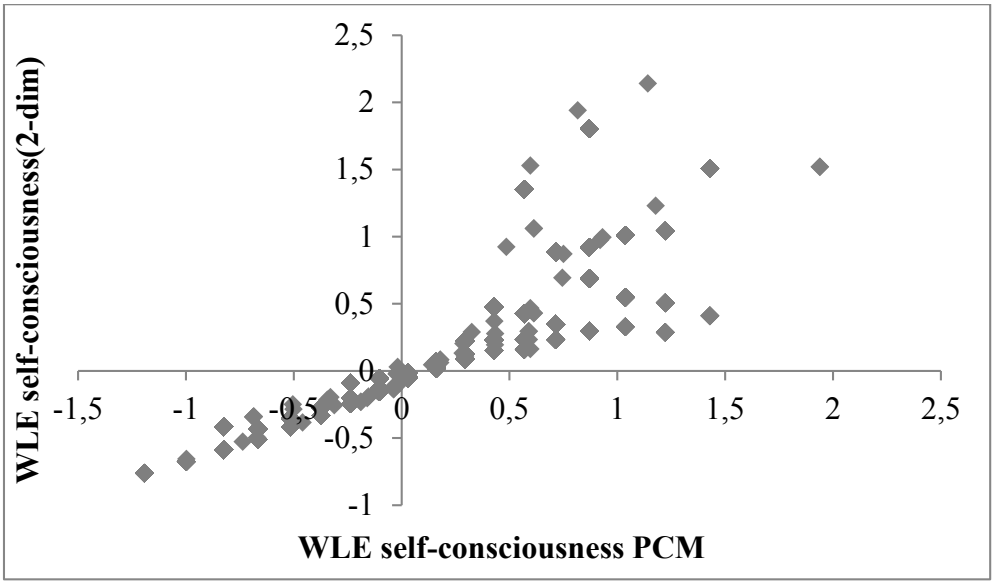

Figure 3. Scatterplot of weighted likelihood estimates (WLE) from the unidimensional partial credit model (PCM) and the two-dimensional model with trait and ERS1 (2-dim) for self-consciousness.

In models that use the same items as the ones used for the trait dimension but model them as separate item sets, trait estimates are not adjusted for ERS (e.g., 3 dim – same modeled as separate in Table 5). The same is the case when items from different (uncorrelated) traits are used to model the ERS dimension (modeling ERS with separate items; e.g., 3 dim – separate and different). Thus, the correction of trait estimates does not universally occur in multidimensional models but is a question of whether trait variance can be separated from response style variance or not. Which is the case is determined by the method of modeling the ERS dimension. As argued above, in models that use the same items to model both traits and response styles, a correction of trait estimates can be observed since response style variance is accounted for in the estimation of trait variance. However, when additional items are used to model response styles, response style variance and trait variance cannot be modeled perfectly

in the estimation of trait variance. In consequence, the more items that do not assess the trait of interest are added to the response style dimension, the less the corrective effect works. In models where separate item sets are used to model traits and response styles trait variance will be confounded with response style variance and thus, trait estimates are not corrected for response style effects. Furthermore, latent correlations between ERS and some NEO-PI-R facets indicate that the ERS assessed by the items on one scale is not necessarily the same as the ERS assessed by the items on another scale. Instead, ERS appears to contain a scale-specific component that does not generalize across scales.

To illustrate the correction of trait estimates using several methods, we additionally applied a corrective method based on regression residuals which can also be used with simple response style indices (e.g., Baumgartner & Steenkamp, 2001). Here, WLEs for the traits are estimated in unidimensional or multidimensional models that do not include response style dimensions. Then, the trait WLEs are regressed on a response style index which in the case of ERS can simply consist of the number of extreme responses a person gave on one trait, several traits, or all traits in the questionnaire. The resulting residuals can be used as alternative trait estimates. The indicators for ERS we applied were based 1) on the eight an-xiety items, 2) the 48 items on all neuroticism facets, and 3) all 240 items in the NEO-PI-R. For low-ERS respondents (less than 10% extreme responses), the distribution of residuals did not differ strongly from the distribution of WLEs (e.g., for ERS based on 240 items: M_{WLE} = 0.08, SD_{WLE} = .70; M_{Res}= -0.01, SD_{WLE} = .67). However, for high-ERS respondents (more than 50% extreme responses), means of residuals were reduced compared to means of WLEs (e.g., for ERS based on 240 items: M_{WLE} = .42, SD_{WLE} = 2.03; M_{res}= 0.13, SD_{res}= 1.95). Thus, taking the trait distributions into account, it appears that trait estimates of high-ERS respondents are adjusted to be less extreme which validates that a correction for ERS takes place.

Table 5

Response Patterns, Sum Scores, and Weighted Likelihood Estimates (WLE) for Two Exemplary Cases

Model	Case	Response pattern N1	Response pattern N2	Sum score N1	Sum score N2	Sum Score ERS1	WLE N1 (SE)	WLE N2 (SE)	WLE ERS1 (SE)
1 dim PCM	8	31413332	22112133	20	15		0.57 (.40)	-0.07 (.37)	
	9700	11144414	00211141	20	10		0.57 (.40)	-0.83 (.43)	
2 dim	8	31413332	22112133	20	15	1 (of 8)	0.44 (.40)		0.93 (.95)
	9700	11144414	00211141	20	10	4 (of 8)	WLEs not estimable		
3 dim	8	31413332	22112133	20	15	1 (of 16)	0.33 (.35)	0.19 (.33)	1.32 (.66)
	9700	11144414	00211141	20	10	7 (of 16)	0.06 (.22)	0.01 (.24)	4.03 (.57)
4 dim	8	31413332	22112133	20	15	1 (of 24)	0.51 (.41)	0.15 (.39)	0.00 (.76)
	9700	11144414	00211141	20	10	12 (of 24)	0.12 (.23)	0.03 (.25)	3.44 (.44)
5 dim	8	31413332	22112133	20	15	1 (of 32)	0.61 (.46)	0.09 (.42)	-1.15 (.91)
	9700	11144414	00211141	20	10	15 (of 32)	0.14 (.25)	-0.03 (.26)	2.97 (.37)
6 dim	8	31413332	22112133	20	15	2 (of 40)	0.98 (.47)	-0.08 (.40)	-0.13 (.43)
	9700	11144414	00211141	20	10	17 (of 40)	0.17 (.27)	-0.05 (.28)	2.57 (.33)
7 dim	8	31413332	22112133	20	15	2 (of 48)	1.01 (.47)	-0.02 (.40)	-0.19 (.41)
	9700	11144414	00211141	20	10	19 (of 48)	0.19 (.27)	-0.07 (.29)	2.40 (.30)
2 dim - same modeled as separate	8	31413332	22112133	20	15	1 (of 8)	0.44 (.40)		0.93 (.95)
	9700	11144414	00211141	20	10	4 (of 8)	0.44 (.40)		2.57 (.71)
3 dim - same modeled as separate	8	31413332	22112133	20	15	1 (of 16)	0.44 (.40)	0.12 (.37)	0.07 (.88)
	9700	11144414	00211141	20	10	7 (of 16)	0.44 (.40)	-0.65 (.43)	2.19 (.51)
2 dim - separate and different	8	31413332	22112133	20	15	0 (of 8)	0.44 (.40)		-1.42 (1.69)
	9700	11144414	00211141	20	10	8 (of 8)	0.45 (.40)		5.28 (1.64)

(continued)

Model	Case	Response pattern N1	Response pattern N2	Sum score N1	Sum score N2	Sum Score ERS1	WLE N1 (SE)	WLE N2 (SE)	WLE ERS1 (SE)
3 dim - separate and different	8	31413332	22112133	20	15	0 (of 16)	0.45 (.40)	0.12 (.37)	-2.02 (1.53)
	9700	11144414	00211141	20	10	15 (of 16)	0.45 (.40)	-0.65 (.43)	4.39 (.91)

Note. N1 = anxiety, N2 = angry hostility, ERS = extreme response style, 1 dim PCM = unidimensional partial credit model, 2 dim to 7 dim = two to seven dimensional partial credit model, 2 dim – same modeled as separate = trait and response style dimensions were modeled using the same items (N1 and N2) treated as separate item sets, 2 dim – separate and different = trait and response style dimensions were modeled using separate item sets where the items used for the response style dimension were from traits that showed low correlations with the traits of interest (warmth and gregariousness).

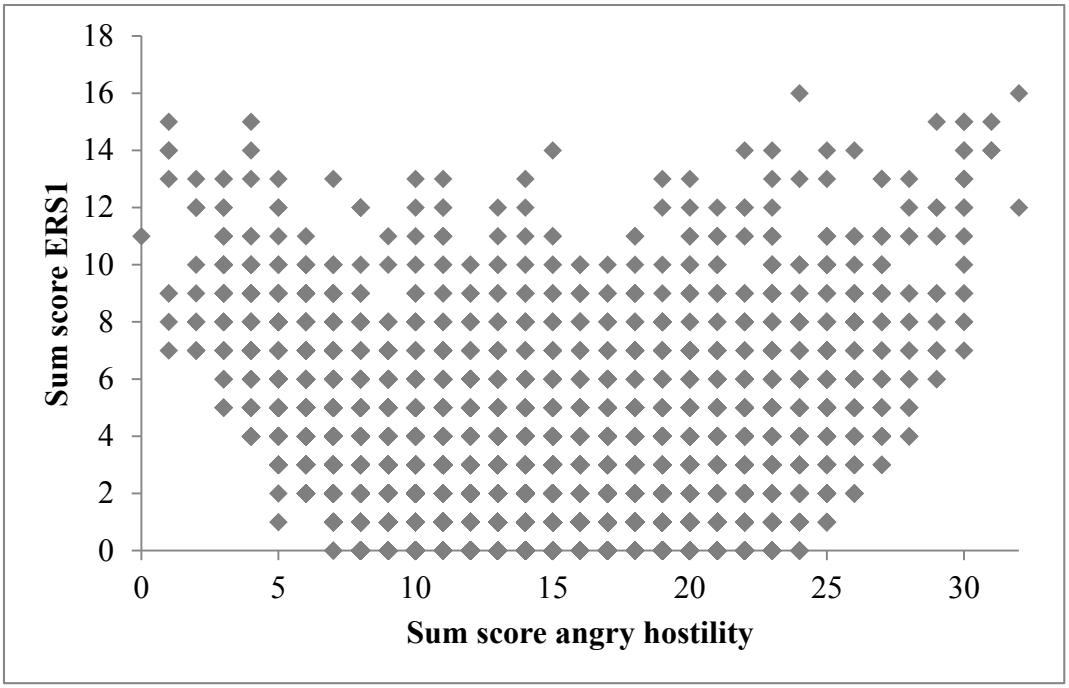

Figure 4a. Scatterplot for sum scores on angry hostility and ERS1.

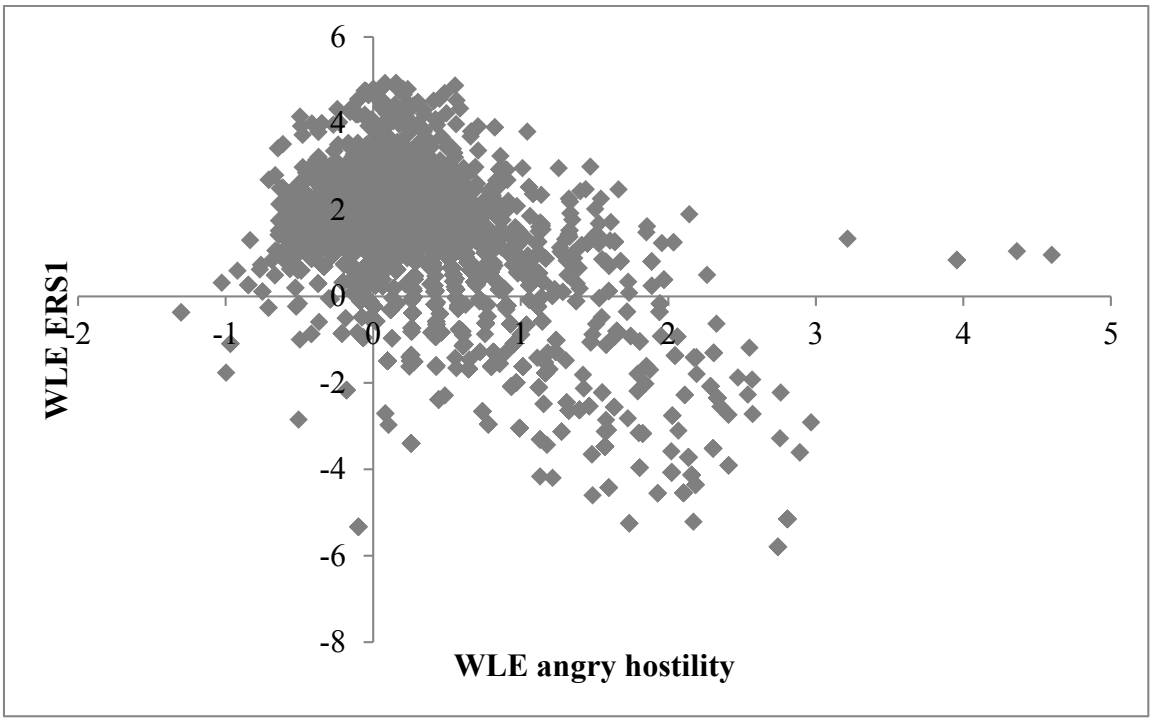

Figure 4b. Scatterplot for weighted likelihood estimates (WLE) on angry hostility and ERS1.

Discussion

In this paper, a multidimensional approach to modeling response styles in the framework of Item Response Theory was taken as opposed to the more common approach of estimating mixed Rasch models to identify latent classes of respondents that differ in their response style (e.g., Austin et al., 2006; Rost et al., 1999). One disadvantage of modeling response styles according to this categorical approach is that when multiple latent classes result in a mixed Rasch model, heterogeneity between classes can be due to several factors, among them differences in response scale usage, differences in the construct being assessed, and differences in the interpretation of the items. To ensure that the same trait is being measured in all latent classes, item difficulties need to be restricted to be the same in all classes (Wetzel et al., 2013). If this model holds against a model with freely estimated item difficulties, differences between latent classes cannot unequivocally be attributed to differences in response scale use. Scales where this is the case should not be used in analyses focused on response styles. Therefore, an advantage of the incorporation of trait and response style dimensions into the same multidimensional model is that all scales can be analyzed. Furthermore, the multidimensional approach allows different response styles to be examined at the same time. This makes it possible to draw model comparisons between unidimensional models and multidimensional models that include different response styles and to investigate relationships between traits and response styles and between different response styles.

In multidimensional models, response styles can be assessed by using the same items as for the trait of interest or by using different items as well. A recommendation given in studies using indices of response styles is to compute the response style index based on items that have low inter-item correlations (e.g., Greenleaf, 1992b) or even a random sample of items from questionnaires assessing heterogeneous traits (Weijters, Geuens, & Schillewaert, 2010b). However, this has the disadvantage that a separate item set has to be administered only for the

purpose of assessing response styles which is often not feasible. Furthermore, ERS indicators from different scales do not necessarily measure the same ERS, implying that ERS contains a scale-specific component. For a different model-based approach to modeling response styles see Meiser and Böckenholt (2011) and Böckenholt (2012).

We compared multidimensional models with the NEO-PI-R traits and different response styles to unidimensional PCMs which assume that only the trait dimension influenced responses. All multidimensional models (irrespective of which response style was modeled) showed a better fit than the unidimensional models. Thus, participants' responses were not only influenced by their trait levels but in addition by at least one response style. This was also supported by our finding that trait estimates from the unidimensional models did not correlate perfectly with trait estimates from two-dimensional models with one trait and one response style. Comparisons between individuals or groups based on sum scores assume that only latent trait levels influence item responses. This assumption neglects the influence of response styles. In our analyses, including an ERS dimension into the model could explain more variance in item responses incremental to the trait dimension than including one of the other response styles (ARS, DRS, and MRS). Thus, ERS appears to be the most important response style. With respect to the two operationalizations of ERS, the clear-cut version which only assigns scores to the two extreme categories (ERS1) worked better. ERS is also the response style that is consistently identified in studies applying mixed Rasch models (e.g., Austin et al., 2006; Rost et al., 1999; Wetzel et al., 2013) whereas ARS or DRS appear not to have been found using this method. However, these studies and our study used samples from countries high on Hofstede's (2001) dimensions of individualism and masculinity (United Kingdom, Germany) where respondents are less likely to employ ARS (Johnson, Kulesa, Cho, & Shavitt, 2005). Thus, it is conceivable that with samples from countries low on collectivism or masculinity one of the other response styles may have greater importance.

When two response styles were modeled simultaneously, the combination of ERS and MRS explained the largest amount of variance. A possible explanation for this finding is that the ERS, ARS, and DRS dimensions are not independent since for all three response styles the extreme categories are scored with 1 in the modeling of response styles. On the other hand, MRS addresses only the middle category and is therefore to some extent independent of the ERS dimension. This is most likely the reason MRS could explain variance in item responses in addition to the trait and ERS.

Another advantage of the approach using multidimensional models is that correlations between traits and response styles or between different response styles can be estimated directly. Unlike the two-stage approach applying a disattenuation formula as commonly implemented in the framework of Classical Test Theory, this direct estimation yields unbiased estimates for the correlations (i.e., without measurement error; Adams et al., 1997; Wang, 1999) in one step. These latent correlations showed that ERS and MRS are mainly trait-independent whereas ARS and DRS show strong associations to several traits. Concerning ERS we only found negative relationships with the agreeableness facets altruism, compliance, and modesty. This finding contradicts Austin et al. (2006) who reported a positive relationship between ERS and both extraversion and conscientiousness in the NEO-FFI. However, ARS and DRS were related to multiple traits. For example, ARS correlated negatively with all neuroticism facets except for impulsiveness for which a small to moderate positive correlation was found. These results coincide with de Jonge and Slaets (2005) who found a significant relationship between positive answers in a questionnaire with empty questions (i.e., only a response scale was presented for each item) and low neuroticism. They also found a relationship between positive answers and high extraversion which could not be confirmed in our study with the exception of the high positive correlation between ARS and the extraversion facet positive emotions.

Furthermore, ARS was positively related to all conscientiousness facets, especially achievement striving (r = .824) while DRS was negatively related to conscientiousness. Future research could aim at investigating in how far the relationship between trait and ARS or DRS depends on the social desirability of the items assessing the trait since, for example for conscientiousness and neuroticism, a relationship to social desirability has been shown (Ones, Viswesvaran, & Reiss, 1996).

Latent correlations between different response styles showed that ARS and DRS were strongly related across the Big Five domains. While the correlation between ARS and DRS was high, it was far from unity and also differed between traits. In fact, it was highest for extraversion (r = .819) and lowest for openness to experience (r = .672). Hence, ARS and DRS show a large amount of overlap, but do not appear to be opposite poles of one dimension. Instead, our findings suggest that it is justified to model them as two separate dimensions. We included moderate agreement and disagreement in our operationalization of ARS and DRS. A high correlation between ARS and DRS when only *strongly agree* and *strongly disagree* are used for coding ARS and DRS might be attributed to ERS dominating the two other response styles (Bachman & O'Malley, 1984). Weijters, Geuens, and Schillewaert (2010a) examined the time-invariant components of four response styles in a study on the stability of individual response styles and also found that ARS was positively associated with DRS. However, Weijters et al.'s finding that ARS was negatively associated with MRS could not be confirmed in our study since our results indicate that ARS and MRS are positively related. ERS does not appear to be highly related to the other response styles though small to moderate correlations with ARS were found.

According to Baumgartner and Steenkamp (2001), using negatively worded items is effective at countering ARS. In our study, this was not the case. We found that the relationship between ARS and traits grew stronger with increasing numbers of negatively worded items.

Thus, the number of negatively worded items on a scale can explain the relationship between ARS and traits to a certain, though minor, extent. On negatively worded items, we would expect respondents with high latent trait levels to use the categories *disagree* and *strongly disagree* more often. When they endorse *agree* and *strongly agree* on negatively worded items, this is characteristic of an ARS. A possible explanation for stronger trait-ARS relationships with increasing numbers of negatively worded items may be that negatively worded items are harder to understand or read less carefully by respondents and in consequence evoke more ARS, though this should be investigated further. The same explanation is plausible for the stronger trait-MRS relationships with rising numbers of negatively worded items. However, the effect of the social desirability of the items on the associations between traits and response styles should be considered here as well.

The most important advantage of modeling traits and response styles simultaneously in multidimensional models is the possibility of obtaining trait estimates that are corrected for response styles. The comparison of respondents using their sum scores is problematic when they differ systematically in their response behavior since these differences will impact sum scores. Hence, substantive trait variance cannot be distinguished from response style variance in the computation of sum scores which means that sum scores are biased when response styles play a role. This problem can be avoided by using Item Response Theory trait estimates such as WLEs derived from a multidimensional model since the WLEs for each dimension reflect the parameters of the other dimensions. The procedure of using multidimensional models to estimate ERS and to obtain trait estimates corrected for ERS was already presented by Bolt and Newton (2011). They reported that estimating ERS based on the items from two traits substantially improved the estimation of ERS and in turn the estimation of trait levels. However, our results indicate that a more differentiated view on the correction of trait estimates in multidimensional models than presented in Bolt and Newton (2011) is necessary. First, the

correction of WLEs only works when the same items are used to model the trait and response style dimensions. If a separate set of items is used to measure the response style, a correction of trait estimates does not take place. In the former case, response style effects are accounted for in the estimation of trait variance. This leads to trait estimates that are corrected for a respondent's response style. In the latter case, trait variance and response style variance are confounded since response style effects are not taken into account in the estimation of trait variance. Consequently, trait estimates are not corrected. Second, in models that use the same items for trait and response style dimensions, the correction has differential effects depending on the number of scales the ERS indicators are used from. The more trait dimensions are modeled – and thus the more items are included in the response style dimension – the less the impact of the corrective effect. This can be explained by the confounding of trait variance with response style variance in each trait dimension. Since the response style is modeled using items from different traits, trait variance cannot be corrected for response style effects in the same manner as when the trait and response style dimensions are modeled using exactly the same items. Furthermore, fluctuations in the ERS indicator occur, depending on which items are used to model the ERS dimension, and ERS correlates with several traits. Thus, ERS cannot be seen as an attribute that exists independently of the items or scales used to assess it, but instead ERS appears to have a scale-specific component that should not be neglected when attempting to correct trait estimates for ERS.

Correcting WLEs post-hoc using regression residuals appears to be a stable method that can also be applied on sum scores (see Baumgartner & Steenkamp, 2001). Future research could compare the correction of trait estimates in multidimensional models to the correction that occurs in mixed Rasch models that differentiate latent classes of response style groups according to the categorical approach (Wetzel et al., 2013). A simulation study that compares different methods for correcting trait estimates or scores for response styles (multidimensional

models, mixed Rasch model, residuals) regarding their ability to recover the true latent trait levels would be helpful in elucidating this issue further.

Limitations of the Study

Limitations of this study include that the multidimensional modeling of response styles was applied to a German sample. Since cross-cultural differences in response styles exist (e.g., Johnson et al., 2005), our results may not generalize to respondents from other cultural backgrounds. Furthermore, the correction of trait estimates in multidimensional models was illustrated using weighted likelihood estimates which are only one of several trait estimates in the framework of Item Response Theory. It can be assumed that results would be similar for other trait estimates such as expected a posteriori or plausible values, though this should be confirmed empirically.

References

Adams, R. J., Wilson, M., & Wang, W.-C. (1997). The Multidimensional Random Coefficients Multinomial Logit Model. *Applied Psychological Measurement*, *21*(1), 1–23. doi:10.1177/0146621697211001

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & B. F. Csaki (Eds.), *Second international symposium on information theory* (pp. 267–281). Budapest, Hungary: Academiai Kiado.

Andersen, E. B. (1985). Estimating latent correlations between repeated testings. *Psychometrika*, *50*(1), 3–16. doi:10.1007/BF02294143

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

Bachmann, J. G., & O'Malley, P. M. (1984). Yea-Saying, Nay-Saying, and Going to Extremes: Black-White Differences in Response Styles. *Public Opinion Quarterly*, (491-509).

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

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443–459. doi:10.1007/BF02293801

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.

Böckenholt, U. (2012). Modeling multiple response processes in judgment and choice. *Psychological Methods.* Advance online publication. doi:10.1037/a0028111

Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: John Wiley & Sons.

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. doi:10.1177/0146621608329891

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

Borkenau, P., & Ostendorf, F. (1993). *NEO-Fünf-Faktoren-Inventar (NEO-FFI) nach Costa und McCrae*. Göttingen, Germany: Hogrefe.

Bozdogan, H. (1987). Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, *52*(3), 345–370. doi:10.1007/BF02294361

Carstensen, C. H. (2000). *Mehrdimensionale Testmodelle mit Anwendungen aus der pädagogisch-psychologischen Diagnostik:* [Multidimensional test models with applications from educational psychological assessment]. Kiel, Germany: IPN.

Cattell, R. B., Cattell, A. K. S., & Cattell, H. E. P. (1993). *16PF fifth edition questionnaire*. Champaign, IL: Institute for Personality and Ability.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. New York, NY: Erlbaum.

Costa, P. T., & McCrae, R. R. (1992). *Revised NEO personality inventory (NEO-PI-R) and NEO Five-Factor inventory (NEO-FFI)*. Odessa, FL: Psychological Assessment Resources.

de Jonge, P., & Slaets, J. (2005). Response sets in self-report data and their associations with personality traits. *European Journal of Psychiatry*, *19*, 209–214.

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

Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika*, *56*(3), 495–515. doi:10.1007/BF02294487

Glas, C. A. W., & Verhelst, N. D. (1995). Tests of fit for polytomous Rasch Models. In G. Fischer & I. W. Molenaar (Eds.), *Rasch-Models. Foundations, Recent Developments, and Applications* (pp. 325–352). New York, NY: Springer-Verlag.

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. doi:10.1086/269326

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

Hofstede, G. (2001). *Culture's consequences*. Thousand Oaks, CA: Sage.

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. doi:10.1177/0022022104272905

Kelderman, H. (1996). Multidimensional Rasch Models for partial-credit scoring. *Applied Psychological Measurement*, *20*(2), 155–168. doi:10.1177/014662169602000205

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*(8), 611–631. doi:10.1177/0146621607312613

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174. doi:10.1007/BF02296272

Meiser, T., & Böckenholt, U. (2011, September). *IRT-Analyse von Traitausprägung und Antwortstilen in Ratingdaten* [IRT analysis of trait levels and response styles in rating scale data], Bamberg, Germany.

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

Ones, D. S., Viswesvaran, C., & Reiss, A. D. (1996). Role of social desirability in personality testing for personnel selection: The red herring. *Journal of Applied Psychology*, *81*(6), 660–679. doi:10.1037/0021-9010.81.6.660

Ostendorf, F., & Angleitner, A. (2004). *NEO-PI-R: NEO-Persönlichkeitsinventar nach Costa und McCrae*. Göttingen, Germany: Hogrefe.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danmarks Paedagogiske Institut.

Rasch, G. (1961). *On general laws and the meaning of measurement in psychology*. Berkeley, CA: University of California Press.

Rossi, P. E., Gilula, Z., & Allenby, G. M. (2001). Overcoming scale usage heterogeneity: A Bayesian hierarchical approach. *Journal of the American Statistical Association*, *96*(1), 20–31. doi:10.1198/016214501750332668

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, *14*(3), 271–282. doi:10.1177/014662169001400305

Rost, J. (1991). A logistic mixture distribution model for polytomous item resonses. *The British Journal for Mathematical and Statistical Psychology*, *44*, 75–92.

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

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. (*Psychometric Monograph* No. 17). Richmond, VA: Psychometric Society. Retrieved from http://www.psychometrika.org/journal/online/MN17.pdf

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464. doi:10.1214/aos/1176344136

Stegelmann, W. (1983). Expanding the Rasch Model to a general model having more than one dimension. *Psychometrika*, *48*(2), 259–267. doi:10.1007/BF02294021

von Davier, M. (2005). *A general diagnostic model applied to language testing data*. *ETS Research Report: No. RR-05-16*. Princeton, NJ: Educational Testing Service.

Wang, W.-C. (1999). Direct estimation of correlations among latent traits within IRT framework. *Methods of Psychological Research Online*, *4*(2), 47-68.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*(3), 427–450. doi:10.1007/BF02294627

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

Weijters, B., Geuens, M., & Schillewaert, N. (2010b). The individual consistency of acquiescence and extreme response style in self-report questionnaires. *Applied Psychological Measurement*, *34*(2), 105–121. doi:10.1177/0146621609338593

Wetzel, E., Böhnke, J. R., Carstensen, C. H., Ziegler, M., & Ostendorf, F. (in press). Do response styles matter? Assessing differential item funtioning for men and women in the NEO-PI-R. *Journal of Individual Differences*.

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

Wilson, M. (1992). The Ordered Partition Model: An extension of the Partial Credit Model. *Applied Psychological Measurement*, *16*(4), 309–325. doi:10.1177/014662169201600401

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. (2007). ConQuest [Computer software]. Camberwell, Australia: Australian Council for Educational Research.

2.5. Appendix E: Manuscript Wetzel & Carstensen (2013c)

Reversed thresholds in the Partial Credit Model – A reason for collapsing categories?

Abstract

When questionnaire data with an ordered polytomous response format are analyzed in the framework of Item Response Theory using the Partial Credit Model, reversed thresholds may occur. This led to the discussion of whether reversed thresholds violate model assumptions and indicate disordering of the response categories. Adams, Wu, and Wilson (2012) show that reversed thresholds are merely a consequence of low frequencies in the categories concerned and that they do not impact the order of the rating scale. This paper applies an empirical approach to elucidate this topic using data from the NEO-PI-R as well as a simulation study. It is shown that categories differentiate between participants with different trait levels despite reversed thresholds and that reversed categories can be analyzed independently of threshold ordering. We show that reversed thresholds often only occur in subgroups of participants. Thus, researchers should think more carefully about collapsing categories due to reversed thresholds.

Key words: partial credit model, threshold parameters, reversed thresholds, ordered rating scales

Introduction

Ordered rating scales are widely used in the assessment of personality, attitudes, and other latent variables. For example, in the NEO-PI-R (Costa & McCrae, 1992), participants respond on a 5 point Likert-type scale with the options *strongly disagree*, *disagree*, *neutral*, *agree*, and *strongly agree*. Another example for an ordered rating scale are the response categories *never*, *sometimes*, *often*, and *always*. With ordered response scales the underlying assumption is that choosing a higher response category implies a higher trait level.

In modeling responses from ordered rating scales according to the Partial Credit Model (PCM; Masters, 1982) a threshold is defined as the point on the latent trait where the response probability for two neighboring response categories is equal. In applications to questionnaire data the order of the thresholds may not correspond to the ordering of the categories. When reversed (or disordered) thresholds occur, a common practice is to collapse the categories that correspond to the reversed thresholds. In many cases, this pertains to the middle category and the next lower category. For example, both, Rost, Carstensen, and von Davier (1999) as well as Austin, Deary, and Egan (2006) combined the categories *neutral* and *disagree* in their Mixed Rasch analyses of the NEO-FFI (Costa & McCrae, 1992). Rost et al. (1999) argued that since the thresholds between *disagree* and *neutral* and between *neutral* and *agree* were reversed, the middle category neutral was chosen less often than would be expected from the trait distribution. They assumed that this indicated that neutral did not measure an intermediate trait level but instead captured a different dimension. Similarly, Nijsten, Sampogna, Chren, and Abeni (2006) reduced a five-category scale to a three-category scale by collapsing categories. Their rationale was to avoid disordered thresholds which they argued would result in illogical response ordering. For examples of studies that retain all response categories despite the occurrence of reversed thresholds see Eid and Rauber (2000) and Zickar, Gibby, and Robie

(2004). Thus, categories are often collapsed to avoid reversed thresholds. This raises the question of whether reversed thresholds are problematic for the ordering of the response categories, justifying this practice, or whether reversed thresholds do not pose a problem and categories therefore should not be collapsed on the basis of reversed thresholds.

In the first part of this paper the Partial Credit Model will be described briefly. We will outline under which circumstances reversed thresholds occur and discuss whether they impact the order of the response categories. An extensive theoretical treatment of the reversed threshold controversy can be found in Adams, Wu, and Wilson (2012). In the second part, empirical examples applying the PCM and its mixture extension to the NEO-PI-R will be reported. Here the trait differences between participants who choose different response categories will be analyzed. Collapsing categories requires the assumption that this is appropriate for the whole sample. We address this topic by exemplarily illustrating that reversed thresholds might occur in subgroups of participants only. In the third part, a simulation study will be presented in which the ability of a five-point scale to discriminate between persons of different trait levels will be compared between several conditions, namely regular response data and response data where two categories were switched. In sum, the aim of this paper is to explore whether the practice of collapsing categories is justified, both from a theoretical viewpoint regarding the measurement model and from an empirical viewpoint regarding the measurement of trait differences between participants.

PART I: The Partial Credit Model

The Measurement Model

The Partial Credit Model (Masters, 1982) is a polytomous item response model which assumes ordered response categories as they exist in partial credit items (*incorrect, partially*

correct, fully correct) or in questionnaires using unidimensional rating scales (e.g., *strongly disagree* to *strongly agree*). Masters' approach was to develop a model in which the dichotomous Rasch Model (RM; Rasch, 1960) is applied to each pair of adjacent categories. It follows that the PCM contains *m* (*m+1* being the number of response categories) location parameters (δ_{ij}), instead of just one location parameter as in the RM. Each location parameter (introduced as thresholds by Andrich, 1978) marks a category intersection (the point on the latent trait where a response in category *x* becomes more likely than in category *x-1*). If the PCM fits the data, separability of item and person parameters exists (Masters, 1982). Hence, the model parameters can be estimated maximizing the conditional likelihood or maximizing a marginal likelihood.

The mathematical model of the PCM (see Equation 1) gives the probability that person *n* with ability θ_n will respond in category *x* ($x = 0, 1, \ldots, m$) of item *i*. The original notation of β for the latent trait (Masters, 1988) was replaced with the customary θ.

$$\pi_{nix} = \frac{\exp \sum_{j=0}^{x} (\theta_n - \delta_{ij})}{\sum_{k=0}^{m} \exp \sum_{j=0}^{k} (\theta_n - \delta_{ij})}, \quad x = 0, m \ . \tag{1}$$

In Equation 1, δ_{ij} is the parameter associated with the transition between two response categories *j*-1 and *j*. The first term in the denominator constitutes an additional constraint that ensures that all π_{nix} will sum up to 1. For notational convenience, it is defined that $\sum_{j=0}^{0}(\theta_n - \delta_{ij}) \equiv 0$ from which follows $\sum_{j=0}^{k}(\theta_n - \delta_{ij}) \equiv \sum_{j=1}^{k}(\theta_n - \delta_{ij})$.

The Partial Credit Model and Threshold Ordering

While the PCM requires ordered response categories, it does not require that the threshold parameters be ordered as well. Masters (1988, p. 23) states: "In the partial credit model

[…] the item parameters δ_{i1}, δ_{i2}, …, δ_{im} govern the transition between adjacent response categories. Order is not incorporated through values of these locally defined parameters, which are in fact free to take any values at all“. When reversed threshold parameters occur in the analysis of questionnaire data it is often concluded that the order of the response categories is violated (Bühner, 2011, p.520). Contrariwise, when the thresholds are ordered, the response categories are assumed to be ordered as well. Thus, sometimes it is argued that categories need to be collapsed in order to avoid reversed thresholds. Reversed thresholds are assumed to indicate that the data cannot be interpreted according to the order of the rating scale but that another dimension may have influenced responses (Rost et al., 1999). However, as demonstrated by Adams et al. (2012), the derivation of the PCM does not posit a connection between the ordering of the threshold parameters and the ordering of the response categories. Furthermore, this line of argument disregards that threshold parameters merely indicate where the likelihoods of neighboring response categories are equal. The ordering or reversal of threshold parameters does not allow any statement about the ordering of the response categories since the ordering of thresholds is dependent on category probabilities (Adams et al., 2012).

Relationship between Category Probabilities and Threshold Ordering

To understand how reversed thresholds occur, it is important to consider the relationship between category probabilities and threshold ordering. The category probability curves in Figures 1a and 1b show the probability of each response category along the trait continuum for two items. These category probabilities are determined by the number of observations in each category (i.e. if more respondents chose a certain category, its category probability will be higher). In these figures, the threshold is the intersection point between two category probability curves (indicated by the perpendicular lines). It marks the transition from one category having a higher response probability than one adjacent response category to the next category

having a higher response probability. For the first item in Figure 1a, each category has a section on the latent trait where it has the highest likelihood of being chosen among all categories. In this case, thresholds are ordered. Note that for the second item in Figure 1b, the middle category (*neutral*) is never, at no point along the latent trait, the most likely category. This is a consequence of the middle category having a low response frequency. The low category probability for *neutral* leads to the second and third thresholds being reversed. Nevertheless, people with trait levels from about -3 to +3 still have a certain probability of choosing this response option. Furthermore, the middle category's curve is still in between the curves for *disagree* and *agree*. Thus, despite reversed thresholds, the order of the category probability distributions along the trait continuum is preserved. In sum, whether threshold parameters will be ordered or not solely depends on the category probabilities which are estimated from the response frequencies for each category. The ordering of the PCM's categories is independent of the ordering of the thresholds and has to be assumed prior to data analysis (Masters, 1988). For a more detailed formal treatment of the distinction between the ordering of the response categories and the ordering of the thresholds see Adams et al. (2012).

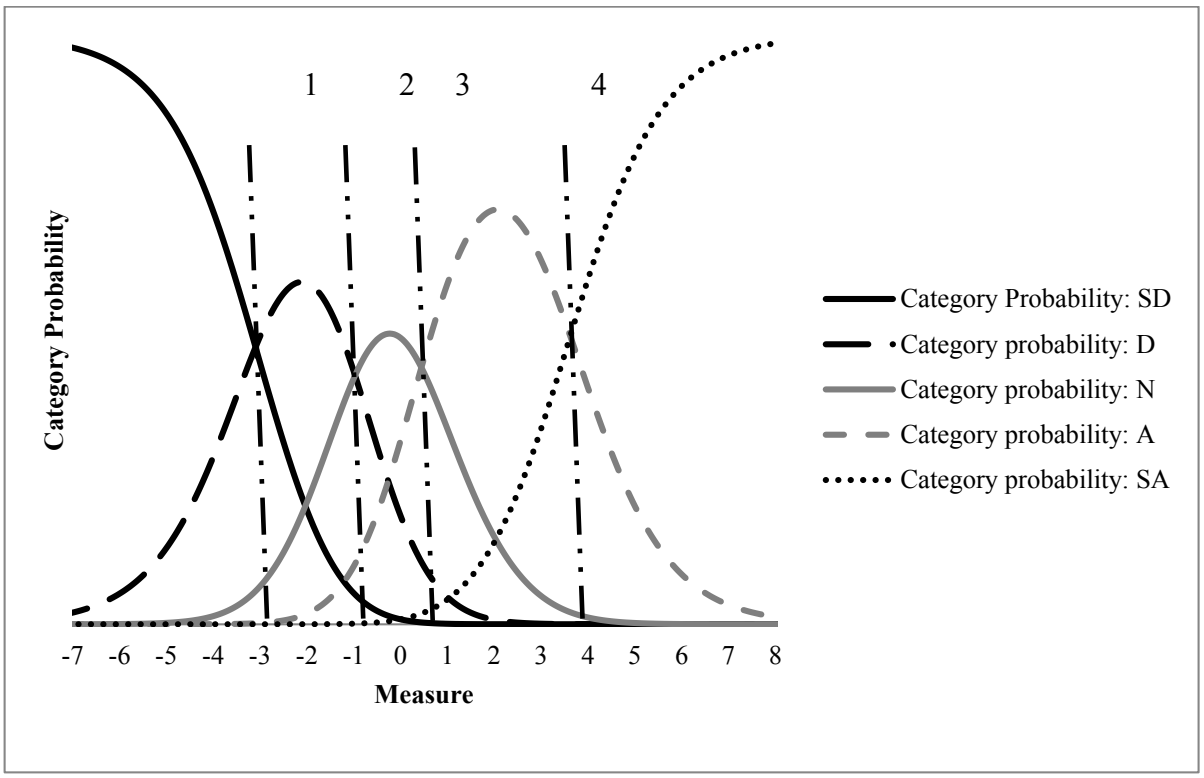

Figure 1a. Category probability curves for item 1 on the Extraversion facet Warmth.

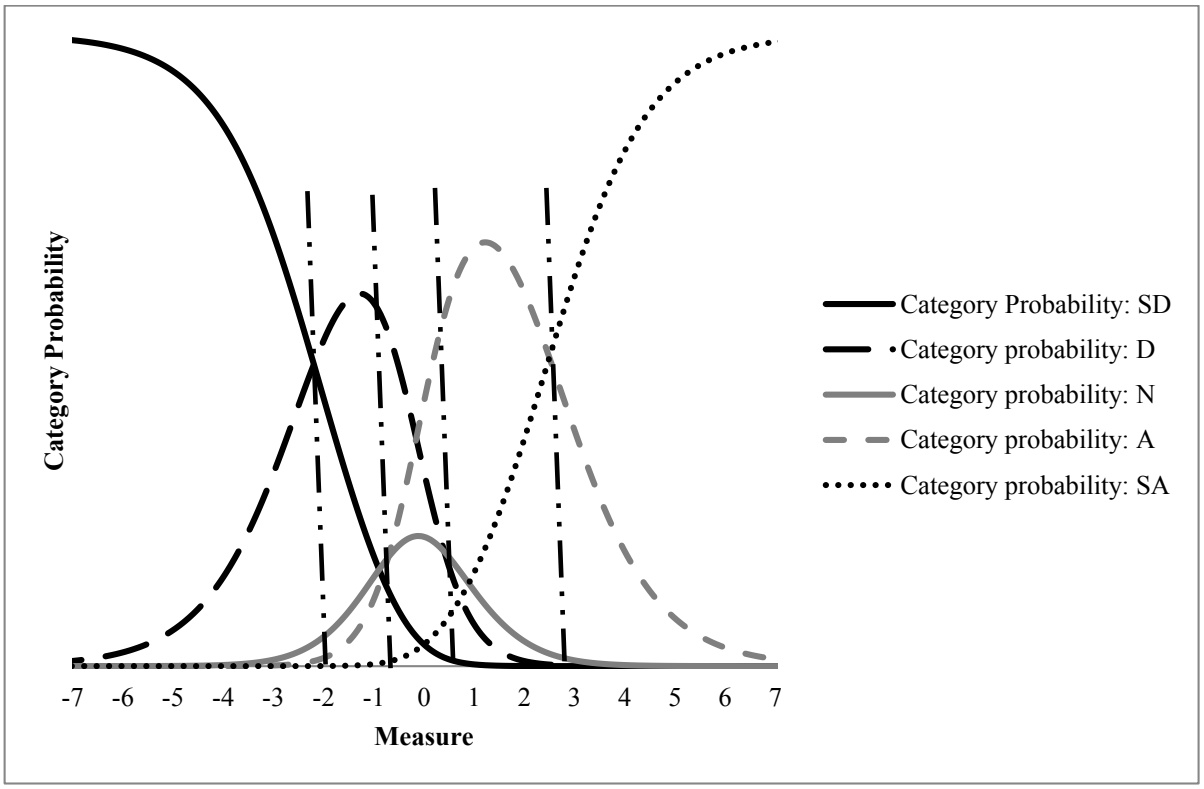

Figure 1b. Category probability curves for item 6 on the Openness to experience facet Openness to actions.

PART II: Empirical Examples from the NEO-PI-R

Using data from the NEO-PI-R, we investigate whether trait estimates derived using the PCM reflect the ordering of the rating scale, i.e. whether persons who choose higher response categories receive higher trait estimates, and whether trait estimates are ordered despite reversed thresholds. Furthermore, differences in trait estimates between categories are analyzed. Collapsing categories requires the assumption that this is appropriate for the complete sample. Using a Mixed Rasch analysis we examine whether reversed thresholds might occur in subgroups of participants only. Following a brief description of the sample and the instrument, the analyses conducted will be depicted, first concerning the trait differences in the Partial Credit Model and next concerning the Mixed Rasch analysis. Then, the results from these analyses will be described.

Method

Sample

The data used here consisted of the German NEO-PI-R's (non-clinical) standardization sample. In total, the dataset contained 11,724 participants (64% women) with a mean age of 29.92 (*SD* = 12.08). Means and standard deviations for the Big Five domains are depicted in Table 1.

Instrument

The German version of the NEO-PI-R (Ostendorf & Angleitner, 2004) was applied. The NEO-PI-R assesses the Big Five personality domains, namely Neuroticism, Extraversion, Openness to experience, Agreeableness, and Conscientiousness. In total, the NEO-PI-R contains 240 items. Each domain consists of 6 subscales (facets) which are assessed by eight items

each. The NEO-PI-R's response scale is a five-point Likert scale ranging from *strongly dis agree* to *strongly agree*. Cronbach's α reliabilities for sum scores on the Big Five domains are reported in Table 1.

Table 1

Cronbach's α Values, Means, and Standard Deviations for the NEO-PI-R Scales

Scale	**Cronbach's α**	***Mean (SD)***
Neuroticism	.93	91.11 (23.57)
Extraversion	.89	110.50 (19.87)
Openness to experience	.89	123.81 (19.36)
Agreeableness	.87	112.63 (16.97)
Conscientiousness	.90	113.90 (20.11)

Analyses

The data were analyzed regarding several aspects. First, trait differences between participants who chose the different response categories were analyzed. Second, a Mixed Rasch analysis was conducted to elucidate how thresholds differ between subgroups of participants.

Trait Differences

We analyzed the data using a PCM in ConQuest (Wu, Adams, Wilson, & Haldane, 2007). ConQuest computes Weighted Likelihood Estimates (WLE; Warm, 1989) as one way of estimating participants' standing on the latent trait. For every item and each response category, an average WLE of the participants that chose that response category is computed. Thus, the trait (WLE) averages for the categories can be compared. The difference in trait averages between categories can indicate whether participants who, for example, chose *neutral*, differ in their trait level from participants who chose, for example, *disagree*. Furthermore, the ordering of the trait averages can be inspected. If the assumption underlying ordered response categories, that persons with higher trait levels choose higher categories, is correct, then this

should be reflected in ordered trait averages from the category *strongly disagree* to *strongly agree*.

Mixed Rasch Analysis

Mixed Rasch Models (MRMs; Rost, 1991) based on the Partial Credit Model were computed using the software WINMIRA (von Davier, 2001). MRMs assume that the RM (or the PCM) holds within latent subpopulations (latent classes) of a sample, but that the model parameters differ across these latent subpopulations (Rost & von Davier, 1995). The mixture generalization of the PCM (MPCM) differs from Master's PCM in that all parameters are specific to each latent class. Otherwise, it has exactly the same properties as the original PCM described above. The appropriate number of classes was determined using the Bayesian Information Criterion (BIC; Schwarz, 1978). Here, results will only be reported for the facet Openness to actions.

Results

Trait Differences

In Table 2, category frequencies, WLE averages for each category, as well as the difference between WLE averages for adjacent categories are depicted for the eight items on the facet Openness to actions. Category frequencies show that *neutral* was chosen by many participants to indicate their standing on the item. In fact, at least for this facet, it was never the least frequent option. The WLE differences between categories range from .23 ($p < .001$; item 3) to .68 logits ($p < .001$; items 5 and 8). In the present context, the difference between *dis agree* and *neutral* is the most interesting. For Openness to actions, it ranges from .30 ($p < .001$; item 6) to .46 logits ($p < .001$; item 5) with a mean of .36 ($SD = 0.05$). Thus, the difference in trait averages for these two categories is comparable to the difference between other categories

and not of a negligible size. In fact, the mean difference in WLE averages between *disagree* and *neutral* computed over all of the NEO-PI-R's 240 items is .42 logits with the 5th percentile at .23 logits and the 95th percentile at .61 logits.

Furthermore, average WLEs from one category to the next increase monotonically for all items in Table 2. Considering the whole NEO-PI-R, there are only eight items where WLE averages are not ordered concerning the categories *strongly disagree* and *disagree* and in two cases additionally concerning *neutral*. This implies that people who chose higher response categories on average have higher trait levels than people who chose lower response categories. WLE averages for the middle category lie between the WLE averages for *disagree* and *agree*. Thus, the middle category *neutral* appears to measure an intermediate trait level.

Mixed Rasch Analysis

The mixture generalization of the PCM was computed for Openness to actions for one to six classes. Openness to actions yielded a four-class solution according to the BIC. Class sizes ranged from 31.82 % to 19.76%. Figure 2 shows the threshold parameters for the first and second latent class of the facet Openness to actions . The four classes can be interpreted as subgroups of participants who differ in their response scale usage. Class 1 (Figure 2a) appears to consist of participants who prefer the options *disagree* and *agree*. In this class, all thresholds are ordered. Class 2 also contains moderate responders but the participants allocated to these classes appear not to use the middle category *neutral* at all as opposed to Class 1 since the second and third thresholds are reversed and widely spaced (Figure 2b). The third class is very similar to the second class. In contrast, participants in Class 4 prefer extreme categories . Importantly, as shown exemplarily for Openness to actions, for most NEO-PI-R facets one class emerged in which thresholds were ordered. For the Openness facets the size of these classes ranges from 11.28% to 46.45%. For a complete treatment of the results of the NEO-

PI-R's analysis using Mixed Rasch Models based on the PCM see Wetzel, Böhnke, Carstensen, Ziegler, and Ostendorf (in press).

Table 2

Item number, Response Category Frequencies, Trait Averages (WLE), and Differences in Trait Averages between Categories, Facet Openness to Actions

Item	Category	Count	Trait avg	Trait avg. SD	Trait avg. difference*
1	SD	345	-0.65	0.71	
	D	2731	-0.27	0.54	0.38
	N	2643	0.08	0.49	0.35
	A	4576	0.45	0.55	0.37
	SA	1410	0.95	0.77	0.5
2	SD	180	-0.72	0.85	
	D	1532	-0.33	0.59	0.39
	N	2972	0	0.53	0.33
	A	5333	0.36	0.57	0.36
	SA	1704	0.82	0.8	0.46
3	SD	822	-0.22	0.76	
	D	5674	0.01	0.59	0.23
	N	2894	0.37	0.57	0.36
	A	2157	0.7	0.7	0.33
	SA	164	1.29	1.14	0.59
4	SD	495	-0.63	0.7	
	D	2600	-0.21	0.53	0.42
	N	2396	0.1	0.52	0.31
	A	4444	0.41	0.57	0.31
4	SA	1771	0.83	0.77	0.42
5	SD	917	-0.49	0.68	
	D	4889	-0.1	0.51	0.39
	N	2825	0.36	0.48	0.46
	A	2695	0.75	0.56	0.39
	SA	381	1.43	0.9	0.68
6	SD	353	-0.56	0.81	
	D	2533	-0.21	0.57	0.35
	N	1850	0.09	0.53	0.3
	A	5800	0.39	0.59	0.3
	SA	1181	0.83	0.85	0.44
7	SD	305	-0.78	0.77	
	D	1638	-0.41	0.49	0.37
	N	2050	0.01	0.52	0.42
	A	5590	0.33	0.53	0.32
	SA	2134	0.82	0.76	0.49
8	SD	565	-0.48	0.71	
	D	4315	-0.14	0.53	0.34
	N	2408	0.22	0.5	0.36
	A	3861	0.59	0.58	0.37
	SA	563	1.27	0.85	0.68

Note. SD = strongly disagree, D = disagree, N = neutral, A = agree, SA = strongly agree, avg. = average.
* All trait average differences are significant at the .001 level.

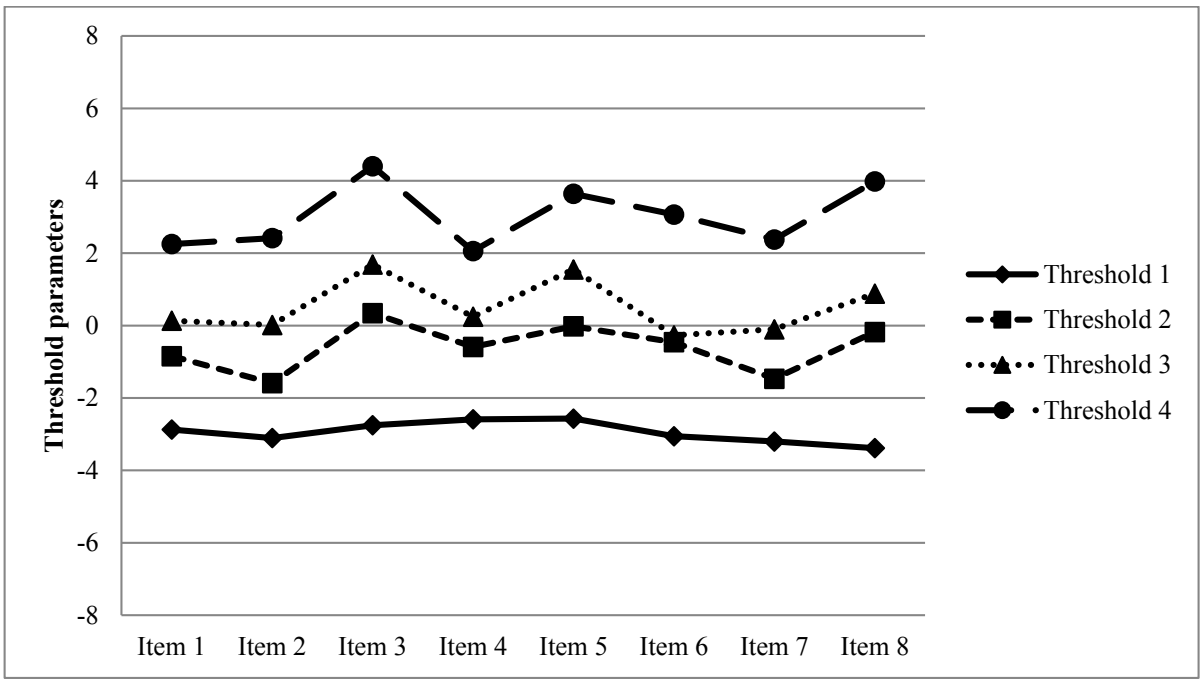

Figure 2a. Threshold parameters for Class 1 (31.82%) on the facet Openness to actions.

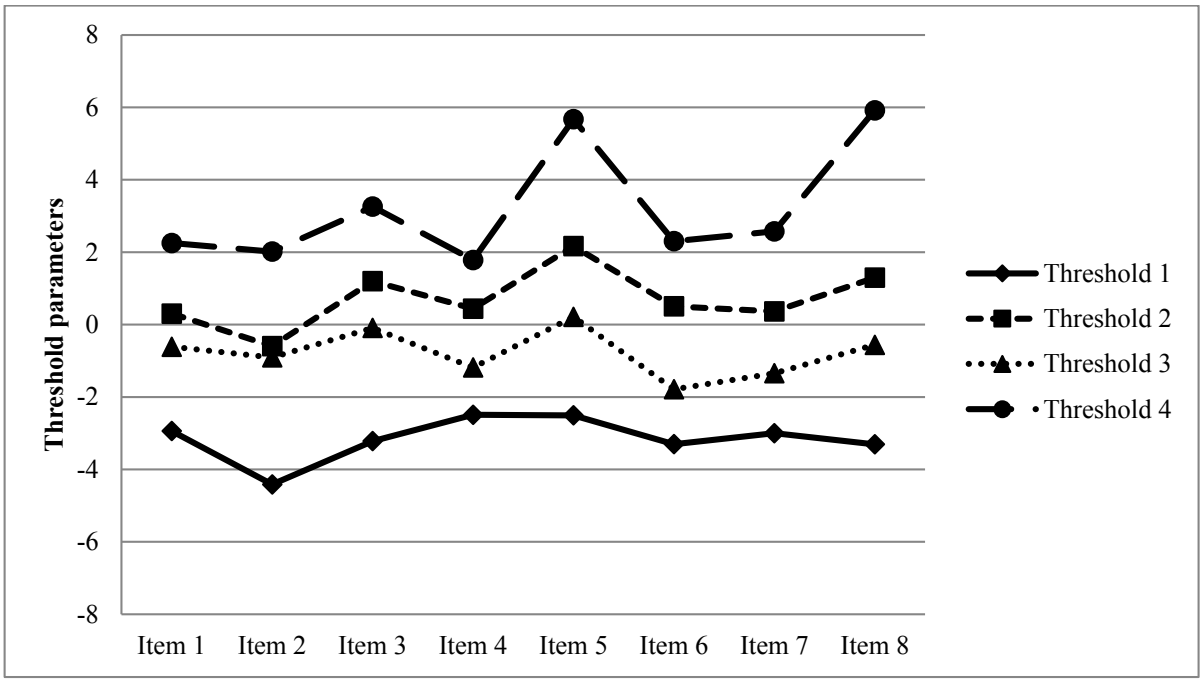

Figure 2b. Threshold parameters for Class 2 (24.22%) on the facet Openness to actions.

Part III: Simulation Study

The aim of the simulation study presented here was to investigate how disordered response data impact parameter estimates of a PCM. The PCM assumes that the response data are ordered and hence, it cannot empirically be tested whether the data are ordered using the PCM. Nevertheless, disordered categories can be detected in the PCM results. We examined how disordered categories influence the distribution of trait estimates and the ability of the response categories to differentiate between participants of different trait levels. We vary the disordering of the categories in two degrees: 1) disordered response categories for one item and 2) disordered response categories for all items in a scale.

Method

Response data based on the PCM were simulated using R 2.12.1 (R Development Core Team, 2010). The data were generated according to the specifications of the NEO-PI-R instrument (eight items with a five-point rating scale). The threshold parameters for the facet Altruism derived from the analyses on the NEO-PI-R standardization sample were used as generating values. In a first step, normally distributed random values for the latent traits were generated for a sample of 5000 subjects. Second, probabilities for a response in each of the five response categories were computed according to the PCM using the generated latent traits and the pre-specified threshold parameters. Then, cumulated probabilities for the response categories were calculated. Next, uniformly distributed random numbers were generated and compared to the cumulated probabilities to determine the responses for the 5000 persons. In total, 100 datasets were generated in this manner.

Lastly, for each replication, two additional data sets with category disordering were created by switching the responses from the second and third category. In one data set responses were switched for the last item of the scale and in the second data set where they were switched for all items. Parameters of the PCM were estimated using ConQuest for all conditions and replications. For each condition, the averages of the trait values (WLE) in each category were inspected regarding their ordering and differences between response categories. Moreover, item discriminations as well as the estimated trait variance from the three conditions were analyzed.

Results

Trait averages were ordered for all items when no items were recoded across all 100 replications. For differing degrees of disorder in the data, trait averages were reversed correspondingly, either only for the last item or for all items. Differences in trait averages between the second and third categories are shown in Table 3. On average, they were larger for the regular dataset compared to the recoded ones, though most notably compared to the dataset where all items had recoded categories. Across 100 replications, the regular datasets yielded differences in trait averages between .63 and .77 logits while the completely recoded datasets yielded WLE average differences between -.12 and -.20 logits. Thus, with switched responses from the second and third categories, trait levels are estimated to be reversed as well as to differ less between categories compared to the original responses.

As depicted in Table 3, discriminations decreased for items where categories are disordered. However, the more items have reversed categories the smaller item discriminations get for the other items in the scale as well. The effect is small if one item has reversed categories and is large if all items have reversed categories. Consistently, the variances of the scales decrease as well, from 1.02 for the regular datasets, to 0.88 when the last item was recoded

and to 0.40 when all items were recoded. Thus, the ability of the items to differentiate between different trait levels was diminished when responses to the second and third categories were switched.

Table 3

Differences in Trait Averages between the Neutral and Disagree Categories and Item Discriminations, Data Generated According to the Three Conditions

	Condition					
	regular		last item recoded		all items recoded	
item	diff. trait averages	discrimi-nation	diff. trait averages	discrimi-nation	diff. trait averages	discrimi-nation
1	.66 (.03)	.71 (.01)	.61 (.03)	.71 (.01)	-.15 (.02)	.60 (.01)
2	.63 (.04)	.73 (.01)	.58 (.03)	.72 (.01)	-.12 (.02)	.63 (.01)
3	.63 (.03)	.73 (.01)	.58 (.03)	.73 (.01)	-.13 (.02)	.64 (.01)
4	.67 (.03)	.70 (.01)	.62 (.03)	.70 (.01)	-.15 (.02)	.59 (.01)
5	.68 (.03)	.70 (.01)	.63 (.02)	.69 (.01)	-.16 (.02)	.58 (.01)
6	.73 (.04)	.64 (.01)	.67 (.04)	.64 (.01)	-.16 (.02)	.57 (.01)
7	.77 (.03)	.64 (.01)	.71 (.03)	.64 (.01)	-.20 (.02)	.50 (.01)
8	.72 (.03)	.67 (.01)	-.35 (.03)	.54 (.01)	-.17 (.02)	.54 (.01)

Note. Diff. = differences.

General Discussion

This paper investigated whether reversed thresholds in the PCM pose a problem in data analysis and whether the practice of collapsing categories might be an appropriate treatment of items with reversed thresholds. Our arguments include a theoretical perspective related to measurement models and an empirical perspective related to the measurement of trait differences. Theoretically, in the framework of the PCM as well as its mixture extensions (Rost, 1991), there is no reason to assume why thresholds would have to be ordered. Adams et al. (2012) show this within several different fundamental derivations of the PCM. Reversed thresholds were shown to be a consequence of (at least) one category not being the most likely category along the whole trait. Thus, whether threshold parameters are ordered or disordered

depends solely on the number of respondents choosing each response category. The occurrence of a reversal does not mean that the order of the response categories is violated since the response categories are still ordered along the trait continuum. Also, considering model fit, items can still function well when reversed thresholds occur (Adams et al., 2012).

Participants who choose different response categories differ strongly in their trait levels as seen in the average WLEs for the five response categories. This was the case for the categories *neutral* and *disagree* in the standardization sample of the NEO-PI-R as well as in the simulation study. Furthermore, as described in the Mixed Rasch analysis, thresholds are often only reversed for a subgroup of participants and not for the whole sample.

When categories are combined, in essence, respondents are treated as if they expressed the same trait level and researchers analyze data as if participants responded to a rating scale with a reduced number of categories. This assumption can hardly be supported by empirical evidence. Considering the large estimated trait level differences between these categories, collapsing categories is not justified.

As shown in the simulation study, in the PCM, the averages of the Weighted Likelihood Estimates for an ordered rating scale are not always ordered. Instead, the PCM estimates the WLEs corresponding to the disordered responses to be reversed. Hence, whether the trait averages per category are ordered along the latent trait measured can be understood as a property of the data. It follows that if the response categories are disordered this can be detected using the participants' trait estimates.

As was evident in the simulation study, when responses to the second and third categories were switched for all items, trait averages were closer together and items discriminated less compared to ordered response data. Also, the trait variance was strongly reduced when the categories of many items were reversed. In sum, reduced item discriminations may hint to

reversed categories and a reduced scale discrimination may be due to reversed categories in a number of items of the scale.

The rationale behind questionnaires using ordered rating scales is usually that more response categories provide more information about the participants' standing on the construct being measured than, for example, a dichotomous *True-False* scale could (Masters, 1988). Considering the large differences between trait averages for the five response categories, this is indeed the case. Collapsing categories counteracts this goal of measuring the latent trait as accurately as possible because it leads to a loss of trait information.

Limitations of this study include that only one type of disorder in the data (namely reversed categories) was simulated. Further research could investigate the impact of different types of disordered data. Moreover, our analyses were empirical examples for questionnaire data similar to the NEO-PI-R. Nevertheless, it was clear in these examples that reversed thresholds do not impair measurement.

In sum, the PCM does not assume ordered threshold parameters and the order of the response categories is preserved even when reversed thresholds occur. Researchers should think more carefully about collapsing categories since valuable trait information is lost.

References

Adams, R. J., Wu, M. L., & Wilson, M. R. (2012). The Rasch Rating Model and the disordered threshold controversy. *Educational and Psychological Measurement*, Advance online publication. doi:10.1177/0013164411432166

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561–573. doi:10.1007/BF02293814 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

Bühner, M. (2011). *Einführung in die Test- und Fragebogenkonstruktion* [Introduction to test and questionnaire construction] (3rd ed.). Munich, Germany: Pearson.

Costa, P. T., & McCrae, R. R. (1992). *Revised NEO personality inventory (NEO-PI-R) and NEO Five-Factor inventory (NEO-FFI)*. Odessa, FL: Psychological Assessment Resources.

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

Masters, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174. doi:10.1007/BF02296272

Masters, G. N. (1988). Measurement models for ordered response categories. In R. Langeheine & J. Rost (Eds.), *Latent trait and latent class models* (pp. 11–29). New York, NY: Plenum Press.

Nijsten, T. E. C., Sampogna, F., Chren, M.-M., & Abeni, D. D. (2006). Testing and reducing Skindex-29 using Rasch analysis: Skindex-17. *Journal of Investigative Dermatology*, *126*, 1244–1250. doi:10.1038/sj.jid.5700212

Ostendorf, F., & Angleitner, A. (2004). *NEO-PI-R: NEO-Persönlichkeitsinventar nach Costa und McCrae*. Göttingen, Germany: Hogrefe.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danmarks Paedagogiske Institut.

Rost, J. (1991). A logistic mixture distribution model for polytomous item responses. *The British Journal for Mathematical and Statistical Psychology*, *44*, 75–92. doi:10.1111/j.2044-8317.1991.tb00951.x

Rost, J., Carstensen, C., & Davier, M. von. (1999). Sind die Big Five Rasch-skalierbar? Eine Reanalyse der NEO-FFI-Normierungsdaten. [Are the Big Five Rasch scalable? A re-analysis of the NEO-FFI standardization data]. *Diagnostica*, *45*(3), 119–127. doi:10.1026//0012-1924.45.3.119

Rost, J., & Davier, M. von. (1995). Mixture distribution Rasch Models. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch Models: Foundations, recent developments, and applications* (pp. 257–268). New York, NY: Springer.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464. doi:10.1214/aos/1176344136

von Davier, M. (2001). WINMIRA 2001 [Computer software]. Kiel: Institute for Science Education.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*(3), 427–450. doi:10.1007/BF02294627

Wetzel, E., Böhnke, J. R., Carstensen, C. H., Ziegler, M., & Ostendorf, F. (in press). Do response styles matter? Assessing differential item funtioning for men and women in the NEO-PI-R. *Journal of Individual Differences*.

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. (2007). ConQuest [Computer software]. Camberwell, Australia: Australian Council for Educational Research.

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

R Development Core Team. (2010). R: A language and environment for statistical computing [Computer software]. R Foundation for Statistical Computing, Vienna. Austria. Retrieved September 2010, from http://www.R-project.org/

2.6. Appendix F: Erklärung

Erklärung

Ich erkläre, das ich die vorgelegte Dissertation selbstständig angefertigt, dabei keine anderen Hilfsmittel als die im Quellen- und Literaturverzeichnis genannten benutzt, alle aus Quellen und Literatur wörtlich oder sinngemäß entnommenen Stellen als solche kenntlich gemacht und auch die Fundstellen einzeln nachgewiesen habe.

Ort, Datum

Unterschrift

2.7. Appendix G: Eigenständiger Anteil an den Manuskripten

Wetzel, E., Böhnke, J. R., Carstensen, C. H., Ziegler, M., & Ostendorf, F. (in press). Do individual response styles matter? Assessing differential item functioning for men and women in the NEO-PI-R. *Journal of Individual Differences*.

- Fragestellung
- Modell/ Methode zu DIF gemeinsam mit J. R. Böhnke; zu MRMs kleiner Anteil
- Datenanalysen
- Erstentwurf und Finalisierung des Manuskripts

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

- Fragestellung
- Modell/Methode goßer Anteil
- Datenanalysen
- Erstentwurf und Finalisierung des Manuskripts

Wetzel, E., & Carstensen, C. H. (2013b). *Multidimensional modeling of response styles*. Manuscript submitted for publication.

- Fragestellung gemeinsam mit C. H. Carstensen
- Modell/Methode gemeinsam mit C.H. Carstensen
- Datenanalysen
- Erstentwurf und Finalisierung des Manuskripts

Wetzel, E., & Carstensen, C. H. (2013c). *Reversed thresholds in the Partial Credit Model – A reason for collapsing categories?* Manuscript submitted for publication.

- Fragestellung gemeinsam mit C. H. Carstensen
- Modell/Methode gemeinsam mit C.H. Carstensen
- Datenanalysen
- Erstentwurf und Finalisierung des Manuskripts

Wetzel, E., & Carstensen, C. H. (2013a). *Linking PISA 2000 and PISA 2009: Implications of instrument design on measurement invariance.* Manuscript submitted for publication.

- Modell/Methode gemeinsam mit C.H. Carstensen
- Datenanalysen
- Erstentwurf und Finalisierung des Manuskripts