A Constrained Factor Mixture Analysis Model for Consistent and Inconsistent Respondents to
Mixed-Worded Scales

Isa Steinmann¹, Rolf Strietholt², & Johan Braeken¹

¹Centre for Educational Measurement, University of Oslo

²International Association for the Evaluation of Educational Achievement (IEA), Hamburg

This manuscript was accepted for publication in *Psychological Methods* on the 9th December, 2020. This document is the prepublication copy of the final manuscript, as accepted for publication.

© 2020, American Psychological Association. This paper is not the copy of record and may not exactly replicate the final, authoritative version of the article. Please do not copy or cite without authors' permission. The final article will be available, upon publication, via its DOI: 10.1037/met0000392

Author Notes

This study was supported by the Gustafsson & Skrondal Visiting Scholarship at the Centre for Educational Measurement at the University of Oslo, by a research grant [FRIPRO-HUMSAM261769] for young research talents of the Norwegian Research Council, and by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant number 765400. We thank Roisin Cronin for copy editing the manuscript.

Correspondence concerning this article should be addressed to Isa Steinmann, Centre for Educational Measurement, University of Oslo, Postboks 1161 Blindern, 0318 Oslo, Norway. E-mail: isa.steinmann@cemo.uio.no

Precursors of this study's ideas and analyses have been presented at the American Educational Research Association (AERA) Annual Meeting and the Tagung Gesellschaft Empirische Bildungsforschung (GEBF) conference.

Abstract

Mixed-worded scales require a more careful reading and answering process than scales with only one type of wording. The present study is about the unintended consequences of using such scales, because typically, not all respondents answer positively and negatively worded items consistently. This population heterogeneity—meaning that there are distinct groups of consistently and inconsistently answering respondents—may arguably underlie wording-related effects in mixed-worded scales. We formulated a constrained factor mixture analysis model that operationalized these two assumed classes of respondents. We investigated five datasets that contained four mixed-worded attitude scales, large-scale samples from three countries (Germany, Australia, and the USA), and two age groups (children and adolescents). The constrained factor mixture analysis model showed estimated parameter patterns in line with theoretical expectations and consistently outperformed its more traditional competitor, confirmatory factor analysis with one global and one orthogonal method factor across all used datasets. We found proportions of between 7% and 20% of respondents belonging to the inconsistent classes. To further substantiate and validate the interpretation of the proposed model, we related class membership to theoretically relevant respondent characteristics such as reading achievement, cognitive skills, or conscientiousness. Further, we undertook an initial exploration of the overlap in inconsistent respondents' class membership across time and across scales within a survey. The paper discusses implications for future research as well as for the use of mixed-worded scales.

Keywords: mixed item wording, factor mixture analysis model, item wording effect, attitude scales, Rosenberg Self-Esteem Scale

Translational Abstract

Many scales combine positively and negatively worded items such as “On the whole, I am satisfied with myself” and “At times I think I am no good at all”. In this case, respondents with high self-esteem have to agree to the positively worded items but disagree to the negatively worded items to give a consistent response. However, previous research suggests that some respondents fail to switch the side on the response scale of mixed worded scales, which has severe consequences for the validity of the survey data. Our study reveals that about one out of five children in primary school and up to one out of 10 adolescents in secondary school fail to switch the side on the response scales. Furthermore, we find that students with low reading ability are overrepresented in the group of students that do not manage to switch the side on the response scale. We replicate these findings using representative large-scale data from Germany, Australia, and the USA and for different attitude scales. We conclude that researchers should be cautious in using such mixed-worded scales in their questionnaires, since a small proportion of respondents seems to have problems with responding adequately to them.

Keywords: mixed item wording, factor mixture analysis model, item wording effect, attitude scales, Rosenberg Self-Esteem Scale

A Constrained Factor Mixture Analysis Model for Consistent and Inconsistent Respondents to Mixed-Worded Scales

Mixed-worded questionnaire scales combine positively and negatively worded items in the same response scale. For instance, in a widely used mixed-worded scale on self-esteem, a respondent who wants to express positive self-esteem has to agree with positively worded items like “I feel that I have a number of good qualities” and has to disagree with negatively worded items like “At times I think I am no good at all” (Rosenberg, 1965). Consistent responses to such mixed-worded scales require the respondent to change the side of the Likert response scale they choose; in one instance, they should check *applies completely*, whereas in the other, they should check *does not apply at all*. The literature uses a variety of terms to describe such phenomena; it uses *positively keyed* or *positively phrased* for positively worded items and *negatively phrased*, *negatively keyed*, *negatively worked*, or *reversed items* for negatively worded items.

The reason for using mixed-worded items is to prevent respondents’ acquiescence, to keep their attention, and to encourage the careful reading of all items before answering. “The basic logic here is that reverse-coded items are like cognitive ‘speed bumps’ that require respondents to engage in more controlled, as opposed to automatic, cognitive processing” (Podsakoff et al., 2003, p. 884; see also Likert, 1974; Nunnally & Bernstein, 1994).

Implementing mixed-worded items in the same scale relies on two assumptions: First, the mixed wording should improve the psychometric properties of a scale by preventing response biases, and second, the scale should measure the same underlying construct(s) that it would have measured if it only contained one type of item wording (see e.g., Benson & Hocevar, 1985; Likert, 1974; Nunnally & Bernstein, 1994).

Evidence for item wording effects in mixed-worded scales. There is a large body of studies showing the unintended consequences of using mixed-worded scales because,

typically, such scales do not just measure the construct(s) of interest; they are also prone to so-called method effects associated with the item wording. When looking at simple item intercorrelation matrixes, researchers have commonly found that positively and negatively worded items are less correlated than expected. In the absence of item wording effects, unidimensional mixed-worded scales should show positive and high intercorrelations between items with the same wording and negative and high intercorrelations between positively and negatively worded items. In other words, respondents who agree with positively worded items should disagree with negatively worded items and vice versa. Yet, empirical studies have found that the intercorrelations between positively and negatively worded items are often more biased towards zero than the intercorrelations between items with the same wording (e.g., Carmines & Zeller, 1974; Dunbar et al., 2000; Marsh, 1986).

Studies that apply a confirmatory factor analytical (CFA) approach offer a complementary perspective. If there were no item wording effects, data derived using unidimensional scales should be well represented in one-dimensional CFA models. However, a large body of literature shows that one-dimensional CFAs have a poor model fit with data from various mixed-worded attitude scales in different age and country samples (e.g., DiStefano & Motl, 2006, 2009; Dunbar et al., 2000; Gnambs & Schroeders, 2017; Lindwall et al., 2012; Magazine et al., 1996; Marsh, 1996; Marsh et al., 2010; Marsh et al., 2013; Quilty et al., 2006; Yang et al., 2012). The model fit typically improves considerably after introducing additional method factors for the item wording or by modeling correlated uniquenesses (*ibid.*). A two-dimensional CFA model—in which both the positively and negatively worded items load on a global substantive factor and the negatively worded items load on an orthogonal method factor—has often been the model of choice (e.g., DiStefano & Motl, 2006, 2009; Gnambs & Schroeders, 2017; Magazine et al., 1996; Marsh, 1986).

In summary, the existence of item wording effects is a well-established finding for a variety of mixed-worded scales. In questionnaires, variance in the responses should, however,

only result from variation in the substantive underlying construct(s), while all remaining variance should be random (random measurement error). Further sources of variation—that is, those that are systematic and independent of the substantive construct(s)—are unintended and may threaten the validity of scales (systematic measurement error). The reason for such unintended variation may lie within the respondent (e.g., carelessness in responding) or the instrument (e.g., complexity of item formulation) (Podsakoff et al., 2003). Therefore, wording-related method effects may undermine the reliability and validity of questionnaire scales (see e.g., Barnette, 2000; Marsh, 1986; Patton et al., 2019; Podsakoff et al., 2003; Schmitt & Stults, 1985; Steedle et al., 2019).

Two explanations for the item wording phenomenon. The literature discusses two central explanations for the observation that unidimensional mixed-worded scales are not well represented by one-dimensional CFA models. In mixed-worded scales, respondents have to notice the changing positive and negative wording while reading the items and then manage to adjust their answers to the response scale accordingly. A double negation may be necessary to express agreement. Noticing the difference in wording or managing to change the side of the answering scale may be too demanding for respondents with low reading or low cognitive abilities. Likewise, the same may apply to a respondent who does not fill out the questionnaire carefully and reads the items only superficially before answering and therefore does not notice the mixed wording. This could be the case for respondents with low commitment or for distracted respondents. Therefore, mixed-worded scales might require higher reading or cognitive skills and a more careful reading and answering process than questionnaire scales with only one type of item wording (see e.g., Benson & Hocevar, 1985; Marsh, 1986; Melnick & Gable, 1990; Schmitt & Stults, 1985; Weems et al., 2003).

Figure 1. Example answers of a consistent (Panel A) versus an inconsistent (Panel B)

respondent to a positively (1) and a negatively (2) worded item of the Rosenberg's self-esteem scale (1965), the most widely used mixed-worded scale in item-wording research, and the implied item intercorrelation patterns.

It seems worth discussing the consequences for the item intercorrelation matrix of the full sample that arises when a subgroup of respondents fails to adjust their responses to the mixed wording. To consistently respond to mixed-worded scales, the respondent has to change the side of the response scale they check when answering a positively worded item as opposed to a negatively worded one. For instance, the example respondent in Panel A in Figure 1 expresses high self-esteem on both items because they agree with a positively worded item and disagree with a negatively worded one. Such consistent responses result in a

negative intercorrelation between positively and negatively worded items. In contrast, an inconsistent response means that the respondent keeps checking the same side of the response scale irrespective of the item wording. The response pattern in Panel B in Figure 1 implies high self-esteem on the positively worded item but low self-esteem on the negatively worded item. Such inconsistent responses lead to positive intercorrelations between positively and negatively worded items.

Therefore, the finding of differential item intercorrelations as a function of item wording (i.e., intercorrelations between positively and negatively worded items that were unexpectedly closer to zero) could be explained by the presence of population heterogeneity, which manifests in the form of a sample made up of both consistent and inconsistent respondents. In a sample of this kind, the positively and negatively worded items would correlate less with each other as the groups' response behaviors would cancel each other out for correlations between mixed-worded items; this would not be the case for same-type items (cf. Figure 1, which contains both types of implied correlation matrices). In the same vein, the presence of the two types of respondents might explain why a one-dimensional CFA model did not fit the data of mixed-worded scales in numerous studies. Two simulation studies support this interpretation as they showed that even small shares of respondents with the described inconsistent response behavior could lead to a misfit of CFA models that reflect the intended data structure (Schmitt & Stults, 1985; Woods, 2006). Similarly, Reise, Kim, Mansolf, and Widaman (2016) showed that the majority of respondents in a sample exhibited response behavior that could be well captured in a one-dimensional CFA, but they found that there was still a minority of model-incompatible respondents. The potential existence of such population heterogeneity in the form of the presence of consistent and inconsistent respondents will be a cornerstone of this study and the subsequent modelling approach.

Empirical evidence for the association between item wording method effects and person characteristics. The key argument of the present study is that the phenomenon of

differential item intercorrelations as a function of the mixed item wording can be explained by the presence of a subgroup of respondents who fail to notice or properly adjust their response behavior to the changing item wording, possibly due to carelessness or a lack of skills. The findings of a number of studies have supported the assumption that item wording effects are linked to person characteristics like reading achievement, cognitive abilities, or self-reported conscientiousness (Dunbar et al., 2000; Gnambs & Schroeders, 2017; Marsh, 1986; Quilty et al., 2006). The hypothesized connection between the item wording effects and person characteristics has been further supported by a longitudinal study that found that these effects are rather stable over time (Motl & DiStefano, 2002). Therefore, the findings suggest that item-wording effects *could* be related to at least some respondent characteristics. However, not all studies found significant relationships between item wording effects and verbal abilities (Greenberger et al., 2003) or attained levels of education (Tomás et al., 2013). It is worth noting that none of the previously discussed studies directly connected their findings to the assumption of population heterogeneity in terms of consistent versus inconsistent respondents to mixed-worded scales.

Empirical evidence for the association between inconsistent responding and person characteristics. Another group of studies did not aim to model the substantive constructs in factor models but investigated persons with consistent versus inconsistent responses. There is evidence that inconsistent respondents who did not switch from one side of the response scale to the other when answering positively versus negatively worded items had lower levels of education and academic achievement than consistent respondents (Melnick & Gable, 1990; Steedle et al., 2019). Furthermore, inconsistent respondents were also more careless than consistent respondents (Kam & Chan, 2018). In this study, careless respondents were identified by their incorrect responses to so-called instructed response items. Such items are scattered throughout the questionnaire and prompt the respondent to check a specific response box if they are reading the items attentively. These studies did not,

however, connect their findings to the body of research that investigated item wording effects in mixed-worded scales.

Conclusions from Previous Research for the Present Study

One set of research papers investigated substantively meaningful latent constructs measured by mixed-worded scales in factor analyses and aimed to partial out the item wording effects, for instance by adding an orthogonal method factor or correlated uniquenesses for the negatively worded items. Such approaches are particularly suitable for partialing out method-related covariance that is unrelated to the substantive constructs (cf. Cole et al., 2007). Such variable-centered approaches assume population homogeneity, since the item wording effect applies to all respondents to a certain degree. In other words, they assume that the item wording effect lies within the instrument and not the respondent (cf. Podsakoff et al., 2003). However, they only enable researchers to investigate the mechanisms behind the occurrence of item wording effects to a limited extent. A method factor is driven by data rather than theory and how a negative or positive score on such a method factor is to be interpreted remains an open question. In the studies where the item wording effect phenomenon is found to relate to person characteristics, it is by implication very difficult to interpret the reasons for and the magnitude and direction of these associations, as well (see e.g., Gnambs & Schroeders, 2017; Reise et al., 2016). Another set of studies had a completely different focus and used mixed-worded scales to identify inconsistent respondents. Such person-centered studies showed that the inconsistent respondents differed from consistent respondents in various ways. Yet, these studies did not aim to measure the substantive constructs underlying the mixed-worded scales.

Our study aimed to combine these two strands of variable- and person-focused research and to simultaneously investigate both substantive constructs and the consistent versus inconsistent respondents to mixed-worded scales. We argue that the phenomenon of differential item intercorrelations in mixed-worded scales is grounded in the existence of two

types of respondents. Some respondents exhibit inconsistent response behavior in relation to mixed-worded scales. Inconsistent respondents do not shift to the opposite side of the response scale when answering positively versus negatively worded items. We discussed two reasons for why respondents may give these inconsistent responses. First, they might fail to understand the item wording or to adjust their response behavior adequately due to low reading or cognitive abilities. Second, they might not notice the mixed wording due to carelessness. In other words, we assumed that the source of the item wording effect lies in the interplay between the respondent (i.e., low reading or cognitive abilities, careless responding) and the instrument (i.e., mixed wording more difficult) (cf. Podsakoff et al., 2003).

The Constrained Factor Mixture Analysis Model Approach to Identifying Consistent and Inconsistent Respondents in the Present Study

The present study aimed at identifying two groups of respondents to unidimensional mixed-worded scales: respondents who answered consistently and respondents who answered inconsistently across positively and negatively worded items. For this purpose, we formulated a constrained factor mixture analysis (FMA) model (e.g., Lubke & Muthén, 2005; Masyn et al., 2010). This model combined a latent continuous factor measuring the construct of interest (i.e., self-esteem in Figure 1) with two latent categorical classes reflecting the type of respondent. We will first introduce the general factor mixture model and then introduce the theoretically motivated constraints that tailored the model to the mixed-worded phenomenon.

Factor Mixture Analysis

In a factor mixture analysis, the multivariate density of the outcome variables $\mathbf{Y}$ is modeled as a weighted sum of component multivariate densities:

$$f(\mathbf{Y}) = \sum_{k=1}^K \pi_k f_k(\mathbf{Y})$$

with the usual restrictions applying to the component weights, $\pi_k > 0 \forall k$ and $\sum_{k=1}^K \pi_k = 1$.

Such a model can also be expressed as a latent class model, with a categorical latent variable C consisting of K classes

$$f(\mathbf{Y}) = \sum_{k=1}^K Pr(C = k) h_k(\mathbf{Y} | C = k).$$

Here, class membership is unobserved but inferred from the observed indicators $\mathbf{Y}$. Within each class, the latter follow a specified distribution $h_k(\mathbf{Y} | C = k)$, which can vary between classes.

Assuming a simple one-factor model in each class, conditional upon class membership, the model for Y_i , the response to item i , can be expressed as

$$(Y_i | C = k) = v_{k,i} + \lambda_{k,i} F_k + \varepsilon_{k,i}$$

where $v_{k,i}$ is the class-specific intercept of item i , $\lambda_{k,i}$ is the class-specific loading of item i on the latent factor F_k , and $\varepsilon_{k,i}$ is the class-specific residual term of the item. Imposing the traditional multivariate normality distributional assumption for the residual terms $\boldsymbol{\varepsilon}_k$, with a zero-mean vector and, a diagonal variance-covariance matrix $\boldsymbol{\Psi}_k$ (i.e., uncorrelated residuals), and a normality distributional assumption for the latent factor, with a class-specific mean α_k and variance Φ_k , this leads to the following conditional distribution within a class for the outcome variables

$$(\mathbf{Y} | C = k) \sim N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

with a mean vector $\boldsymbol{\mu}_k = \mathbf{v}_k + \Lambda_k \alpha_k$ (Λ_k representing the class-specific vector of item loadings $\lambda_{k,i}$) and a variance-covariance matrix $\boldsymbol{\Sigma}_k = \Lambda_k \Phi_k \Lambda_k^T + \boldsymbol{\Psi}_k$. The latter distributional form is easily recognizable as the common factor model but now has class-specific parameters.

This general model is, however, not identified and requires identification constraints across classes just like in the case of multiple group analysis across groups (e.g., Meredith, 1993). In its current state, the model would also remain merely exploratory and it is unclear

what the latent classes would stand for. This is especially tricky, as these types of mixture models can either pick up on genuine population heterogeneity and subgroups in the data or on incompatibilities of the model and the data in the one-class model, resulting in the estimation of extra classes to accommodate these without any substantive interpretation (see Bauer & Curran, 2004). In the next subsection, we will elaborate on how the model can be further constrained to reflect our theoretical considerations with respect to consistent and inconsistent respondents to mixed-worded scales.

The Constrained FMA Model

In previous sections, we noted that this study considers two groups: a majority group of consistent respondents and a smaller group of inconsistent respondents, who respond normally to positively worded items but do not properly switch sides on the response scale when answering a negatively worded item. Hence, we have $K = 2$ latent classes, representing the consistent and inconsistent respondents, respectively. Each of these will have their own measurement model, with the latent factor defined by class-invariant item parameters for positive items and specifically constrained item parameters for negative items, i.e. class-invariant item parameters with opposite factor loading signs. Given our theoretical assumptions, the class-invariance of the positively worded items and the polar opposite class invariance of the negatively worded items aim to ensure that the latent factor has the same substantive meaning (i.e., self-esteem in the example in Figure 2) in both latent classes. The parameterization is summarized in the form of the class-specific path diagrams in Figure 2.

Positively worded items. We assume that both classes—that is, the consistent and inconsistent respondents—respond similarly to positively worded items. This translates into, in factor analytical terms, a strong measurement-invariance constraint across classes for the positively worded items. We set both the intercepts and factor loadings of the positively worded items to be equal across classes (i.e., $\lambda_{1,i}^+ = \lambda_{2,i}^+$ and $v_{1,i}^+ = v_{2,i}^+$). To set the poles of the

latent scale, we also constrained the factor loadings of the positively worded items to be positive (i.e., $\lambda_{k,i}^+ > 0$).

Negatively worded items. Because we are using the original responses for all items and not reverse-coded ones for the negatively worded items, we would expect positive loadings on positive items (i.e., greater self-esteem implies more agreement) and negative loadings on negative items (i.e., greater self-esteem implies less agreement) for the consistent class. Hence, for the consistent class, the factor loadings on negatively worded items were constrained to be negative (i.e., $\lambda_{k,i}^- < 0$). In contrast, because the inconsistent respondents do not switch to the proper side of the response scale when marking negatively worded items, greater self-esteem would now (wrongly) imply more agreement. We translated this into a strong model constraint by switching the signs of the factor loadings between the two classes for negatively worded items. Besides this difference in direction, the items function as equally good indicators across classes (i.e., $\lambda_{1,i}^- = -\lambda_{2,i}^-$). The intercepts of the negatively worded items ($v_{1,i}^-$ and $v_{2,i}^-$) were estimated freely in both classes.

Identification constraints. The proposed model has the feature of partial strong invariance, and, as such, we set the metric of the latent factor by standardizing the factor in the class of consistent respondents, who functioned as the reference class in the model. In the inconsistent class, we freely estimated the factor's mean (α_2) and variance (ϕ_2). Although identification practices similar to those used in the multigroup case apply here, the corresponding identification constraints are necessary but not sufficient for identification. The fact that group membership is unobserved does lead to additional complications. In practice, when the size of one or more classes is small relative to the number of class specific parameters, empirical underidentification of the mixture model is likely to arise (Lubke & Luningham, 2017). This will be the case here, as the inconsistent class will be a typical minority class. To prevent empirical underidentification issues, we applied a homoscedasticity

assumption across classes to the residual variances in the finite mixture models (i.e., $\psi_{1,i} = \psi_{2,i}$).¹

Model implications. The combination of the class-invariant constraints for the positively worded items and the reflected-loadings constraints for the negatively worded items will result in the consistent class showing negative intercorrelations between mixed-worded items, whereas the inconsistent class shows positive intercorrelations but similar positive correlation patterns among items with the same wording (see also Figure 1).

Figure 2. Constrained factor mixture analysis (FMA) model for the self-esteem scale, which includes five positively worded (Pos1–Pos5) and five negatively worded items (Neg1–Neg5).

¹ We also estimated less constrained FMA models in which the residual variances of the indicators were estimated freely between and within classes. These models did not converge for three out of five datasets (see dataset overview in Table 1 below). The two models that converged showed the highest intercorrelations between items with the same wording in the raw data. This anecdotal observation could indicate that in case of stronger factor structures, the residual variance constraints can be more relaxed than in our constrained FMA model. This could be one relevant direction for further research on empirical identification issues in factor mixture models in general.

Expected findings of the constrained FMA analyses. We generally expected the class of inconsistent respondents to be smaller than the class of consistent ones (cf. Marsh, 1986; Reise et al., 2016; Schmitt & Stults, 1985; Woods, 2006). Since the “distinctive characteristic of self-attitudes [...] is that everyone is motivated to hold the same attitude toward the object, namely, a positive attitude” (Rosenberg, 1965, p. 9), we expected the responses to be skewed towards positive attitudes. Specifically, this tendency towards positive attitudes should be reflected in the intercepts of the positively worded items, which should indicate positive attitudes on average. In the class of consistent respondents, we expected the intercepts of the negatively worded items to be located at the other side of the scale than those of the positively worded items, since this pattern would consistently indicate positive attitudes. In the class of inconsistent respondents, however, the intercepts of the negatively worded items should be similar to those of the positively worded items and therefore express an inconsistently negative attitude. In the model, the item intercorrelations were constrained to be positive for items with the same wording in both classes; they were constrained to be negative for mixed-worded items in the consistent class and positive for mixed-worded items in the inconsistent class. However, we did not expect the magnitude of the intercorrelations to vary much between items with the same or mixed wording, since the less pronounced intercorrelations of mixed-worded items in the full population should only arise due to the existence of the two subpopulations of consistent and inconsistent respondents.

Further validity analyses. We furthermore aimed at validating the two identified latent classes of respondents. We hypothesized that the inconsistent response behavior would be more likely exhibited by persons with lower reading or cognitive abilities and by careless respondents. We therefore investigated the association between latent class membership and reading achievement, cognitive skills, and self-reported conscientiousness. If inconsistent responding related to person characteristics, it should be stable over time and across different mixed-worded scales. We therefore also explored the overlap between membership of the

inconsistent latent classes over time and across two mixed-worded scales in the same questionnaire.

Bridges to previous research. To establish a connection to previous research, we ran two alternative CFA models in the main and further analyses. As a baseline model, we specified a one-dimensional CFA with a single global factor. This represented the intended latent structure of unidimensional mixed-worded scales. The second CFA was a two-dimensional model in which the negatively worded items also loaded on an orthogonal method factor. Such models have been used frequently to partial out item wording effects in previous research.

Methods

In the present study, we investigated four mixed-worded questionnaire scales in three different countries (Germany, Australia, and the USA) and two age groups (children and adolescents) in order to explore the generalizability of our findings (see overview in Table 1). All mixed-worded scales included a balanced or almost balanced number of positively and negatively worded items with the same Likert response scales.

Table 1

Overview of the Five Datasets

No. data- set	Scale	No. pos. items	No. neg. items	Study	Year	Country	Age sample	<i>n</i>	No. schools
1	self-esteem	5	5	NEPS	2010/11	Germany	grade 5	4,989	203
2	self-esteem	5	5	NEPS	2014/15	Germany	grade 9	4,791	238
3	math. self-con.	4	3	PIRLS/TIMSS	2011	Australia	grade 4	5,943	280
4	read. self-con.	4	3	PIRLS/TIMSS	2011	Australia	grade 4	5,943	280
5	read. self-con.	3	3	PISA	2018	USA	15-year olds	4,799	163

Note. NEPS is short for National Educational Panel Study, PIRLS/TIMSS stands for the joint assessment of PIRLS (Progress in International Reading Literacy Study) and TIMSS (Trends in International Mathematics and Science Study), and PISA is short for Programme for International Student Assessment.

Data

Germany. We used two samples from the National Educational Panel Study (NEPS), which is a nationally representative longitudinal study conducted in Germany that contains various tests and questionnaires (Blossfeld et al., 2011).² We included a sample of fifth-graders who participated in 2010/11 and a sample of ninth-graders who participated in 2014/15. We excluded students (a) from an additional oversample of immigrant students, (b) from a special-needs-school sample, and (c) from a survey of individually traced students. Further, we excluded those (d) who did not participate in the main survey. Our samples comprised $n = 4,989$ fifth-graders and $n = 4,791$ ninth-graders. A subsample of $n = 3,032$ students of the NEPS grade 5 sample were also part of the grade 9 sample.

Australia. We used a sample of $n = 5,943$ fourth-graders from Australia, who participated in the joint assessment of the Progress in International Reading Literacy Study (PIRLS) and Trends in International Mathematics and Science Study (TIMSS) in 2011 (Foy, 2013). PIRLS and TIMSS are international large-scale assessment studies that conduct tests and questionnaires in representative samples.

USA. We used a sample of $n = 4,799$ fifteen-year-old students from the USA who were surveyed in the Programme for International Student Assessment (PISA) in 2018 (OECD, 2019). We excluded students who participated in the simplified assessment, the so-called *une heure* test. PISA is an international large-scale assessment that examines fifteen-year-olds from different grades by means of tests and computer-based questionnaires. The adolescents in the U.S. sample in PISA 2018 were enrolled in grades 8 to 12.

² This paper uses data from the National Educational Panel Study (NEPS): Starting Cohort Grade 5, doi:10.5157/NEPS:SC3:9.0.0. From 2008 to 2013, NEPS data was collected as part of the Framework Program for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research (BMBF). As of 2014, NEPS is carried out by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg in cooperation with a nationwide network.

Mixed-Worded Scales

Rosenberg's self-esteem scale in NEPS. NEPS administered a mixed-worded self-esteem scale in the student questionnaires in grades 5 and 9. This scale was developed by Rosenberg (1965) as a 10-item scale to measure a “favourable or unfavourable attitude toward oneself” (p. 18) and is the most widely used scale to investigate item wording effects in mixed-worded scales. NEPS used a German version with five positively and five negatively worded items and five Likert response categories (see German item wording in Appendix 1.1). The English versions of the items are depicted—in the same order as in the NEPS survey—in Table 2. In the NEPS grade 5 sample, Cronbach's alpha was .792; in the grade 9 sample, it was .896.

Mathematics self-concept scale in PIRLS/TIMSS. This scale comprised four positively and three negatively worded items with a four-category Likert response scale. It was assessed in the TIMSS-related part of the joint student questionnaire of the PIRLS/TIMSS assessment in 2011 (see item wording in Appendix 1.2). Mathematics self-concept is an important aspect of academic attitudes. Students who have a strong mathematics self-concept and thus confidence in their ability to succeed are expected to be more motivated to learn mathematics; in consequence, they should engage more and improve their mathematics skills (Mullis, Martin, Ruddock et al., 2009). Cronbach's alpha was .858 in this sample.

Reading self-concept scale in PIRLS/TIMSS. This scale consisted of four positively and three negatively worded items with a four-category Likert response scale (see item wording in Appendix 1.3). It was assessed in the PIRLS-related part of the joint student questionnaire in 2011 and positioned before the mathematics self-concept scale in the questionnaire. Reading self-concept is considered an important aspect of academic attitudes, since students who have a strong reading self-concept and therefore expect to succeed are expected to be more motivated to read; they should consequently read more often for pleasure

and become better readers (Mullis, Martin, Kennedy et al., 2009). Cronbach's alpha was .719 in this sample.

Reading self-concept scale in PISA. The reading self-concept scale in PISA 2018 comprised three positively and three negatively worded items with a four-step Likert response scale in the computer-based student questionnaire (see item wording in Appendix 1.4). Cronbach's alpha was .835 in this sample.

Models

We estimated three models in order to contextualize and investigate the research questions. In line with the outlined central models in previous research, we first ran a one-dimensional CFA in which the positively and negatively worded scale items were indicators of one global latent factor. Given that the mixed-worded scales were designed to measure a one-dimensional construct, this model evaluated the fit of the scales' intended latent structures. Second, we ran a two-dimensional CFA model with both positively and negatively worded items loading on a global factor and with negatively worded items additionally loading on an orthogonal method factor. This model reflected commonly favored models in previous research. The constrained FMA was the third and most important model (see Figure 2). It aimed at identifying two latent classes of consistent and inconsistent respondents. With the exception of the number of positively and negatively worded items, the models of these three analyses were identical for all five datasets (see dataset overview in Table 1).

For all models, the Likert item indicators were regarded as continuous. The skewness and kurtosis statistics of the items in the mixed-worded scales (see Table 2 and Appendix 1.1-1.4) indicated a tolerable deviation from the distribution assumptions for modeling continuous indicators (Curran et al., 1996). Using the items as categorical indicators would have led to a model that would have been closer to the observed data as it would have modeled the occurrence of different responses in more detail. However, our approach just averages over the categories as we do not consider the additional category distinction being crucial for the

study at hand. We acknowledge that all models are wrong (Box, 1976), but we consider ours useful for identifying the latent classes of consistent and inconsistent respondents given the theoretical assumptions. Modeling the indicators as continuous furthermore links our study directly to the previous research in this field. We used the original scores of the positively and negatively worded items (i.e., negatively worded items were not reverse coded).

Validation

Our constrained FMA aimed to identify two latent classes of consistent and inconsistent respondents. In all latent variable models—and especially in mixture-type models—there is, however, a risk of identifying spurious latent factors, classes, or relationships (see e.g., Bauer & Curran, 2004; Courvoisier et al., 2007). To further validate the classes, we investigated whether class membership was indeed associated with theoretically relevant covariates. To avoid the possibility that the inclusion of the covariate would distort the class definition (see e.g., Vermunt, 2010), we used a three-step approach instead of a one-step joint modeling approach. This multistep approach works from the original FMA model to regress the posterior most-likely class membership variable as an outcome on the covariate predictor of interest, while taking misclassification errors in the former into account; this approach has been shown to work well assuming fair class separation (i.e., high entropy) (Asparouhov & Muthén, 2014). In these validity analyses, we *z*-standardized the original covariate scores to facilitate the interpretability of the findings.

As discussed above, inconsistent responses to positively and negatively worded items can be assumed to relate to lower reading and cognitive abilities or to lower conscientiousness scores in personality inventories. NEPS, PIRLS/TIMSS, and PISA all include reading achievement scores from standardized tests (see Gehrler et al., 2012; Mullis, Martin, Kennedy et al., 2009; OECD, 2019).³ In the NEPS grade 5 sample, cognitive abilities were assessed in

³ In the analyses based on NEPS, the reading scores are weighted maximum likelihood estimates. With regard to PIRLS/TIMSS and PISA data, we used the first out of five or ten available plausible values.

two tests. Cognitive reasoning was assessed in a matrices test and perceptual speed was measured in a picture-symbol test (see Haberkorn & Pohl, 2013). In the NEPS grade 9 sample, a two-item conscientiousness scale was assessed in the student questionnaire. The conscientiousness scale was derived as the mean of two items (“I am thorough” and the reverse-scored “I am easy-going and tend to be a bit lazy”) with a Likert scale that ranged from 1 (*does not apply at all*) to 5 (*applies completely*). Cognitive skills or conscientiousness were not assessed in the other samples.

The selection of samples and scales furthermore allowed us to undertake an initial tentative investigation of an overlap between inconsistent respondents over time and across scales in a survey. We used the NEPS data to study the overlap in the independent FMA classifications from grade 5 to grade 9. We used PIRLS/TIMSS data to analyze the overlap in class membership based on the mathematics and reading self-concept scales, which were administered in the same questionnaire. Based on the FMA results, we assigned students to the most likely class based on the estimated highest posterior probability for each individual and studied cross tables of class membership. Although it was less easy to predict what would happen in this case, we expected some overlap in inconsistent respondents if the class membership was indeed associated with stable student characteristics.

As a last validation of the identified classes of consistent and inconsistent respondents, we explored item response patterns for the working example of the NEPS grade 5 dataset. Prototypically consistent respectively inconsistent response patterns, as identified by simple recoding of perfectly consistent and inconsistent response patterns, should lead to a classification as consistent respectively inconsistent respondents in the constrained FMA.

Statistical Analyses

All CFA and FMA analyses were run in *Mplus* version 8 (Muthén & Muthén, 1998-2017) and further processed with the R package *MplusAutomation* version 0.7-3 (Hallquist et al., 2018). Appendices 2.1 to 2.4 depict the *Mplus* syntaxes for all analyses for the first

dataset, the self-esteem scale in NEPS grade 5, for exemplary purposes. We used full-information maximum likelihood (FIML) estimation with robust standard errors (MLR) and the expectation maximization acceleration (EMA) optimization algorithm. We generated 5,000 initial stage random sets of starting values, and used 500 final stage optimizations and 10 initial stage iterations.

Since we used stratified large-scale data, we weighted all analyses. Specifically, we used weights for students from the main sample who participated in the respective measurement points in NEPS (w_{t1} in grade 5 and w_{t5} in grade 9) and student weights for the national samples in PIRLS/TIMSS ($HOUWGT$) and PISA ($SENWGT$). The robust Huber-White sandwich estimator method was used to account for the fact that students were nested in schools in all datasets (see Table 1) (Muthén & Satorra, 1995).

Results

Descriptive Statistics

When looking at the first dataset, the self-esteem scale in NEPS grade 5, we saw that the positively worded items showed higher means than the negatively worded items (see Table 2). In addition, the positively worded items tended to be negatively skewed and vice versa. Thus, the respondents on average tended to agree with positively worded items and disagree with negatively worded items; they expressed high self-esteem on average. Items with the same wording correlated positively with each other (i.e., average item intercorrelation $\bar{r} = .362$ and $\bar{r} = .406$, for positively and negatively worded items, respectively). The item intercorrelations between the positively and negatively worded items were negative but closer to zero ($\bar{r} = -.193$) than the intercorrelations of items with the same wording. This pattern conforms to findings from previous studies, as described above. The other descriptive analyses of the NEPS grade 9, PIRLS/TIMSS, and PISA datasets showed the same pattern (see Appendix 1.1 to 1.4).

Table 2

Item Wording, Descriptive Statistics, and Item Intercorrelations of the Self-Esteem Scale in NEPS Grade 5 in Germany in Original Sequence (Top) and Ordered by Item Type (Bottom)

Items	To what extent do the following statements apply to you?	<i>M</i>	<i>SD</i>	Skewness	Excess kurtosis	Missing values
Pos1	a) On the whole, I am satisfied with myself.	4.29	0.86	-1.09	0.85	2%
Neg1	b) At times I think I am no good at all.	2.45	1.31	0.44	-0.98	4%
Pos2	c) I feel that I have a number of good qualities.	4.08	0.90	-0.78	0.21	3%
Pos3	d) I am able to do things as well as most other people.	4.01	0.96	-0.79	0.15	3%
Neg2	e) I feel I do not have much to be proud of.	2.19	1.31	0.85	-0.47	3%
Neg3	f) I certainly feel useless at times.	1.84	1.17	1.32	0.75	3%
Pos4	g) I feel that I am a person of worth, at least on an equal plane with others.	3.89	1.24	-0.92	-0.17	4%
Neg4	h) I wish I could have more respect for myself.	2.74	1.37	0.18	-1.16	6%
Neg5	i) All in all, I am inclined to feel that I am a failure.	1.80	1.15	1.40	0.98	4%
Pos5	j) I take a positive attitude toward myself.	4.03	1.06	-0.98	0.36	4%

Item intercorrelations										
	Pos1	Pos2	Pos3	Pos4	Pos5	Neg1	Neg2	Neg3	Neg4	Neg5
Pos1	1									
Pos2	.398	1								
Pos3	.351	.451	1							
Pos4	.325	.298	.281	1						
Pos5	.444	.384	.343	.341	1					
Neg1	-.251	-.129	-.147	-.175	-.240	1				
Neg2	-.216	-.157	-.150	-.194	-.224	.338	1			
Neg3	-.310	-.220	-.235	-.214	-.324	.510	.443	1		
Neg4	-.119	-.063	-.089	-.042	-.072	.313	.321	.326	1	
Neg5	-.301	-.216	-.238	-.194	-.303	.449	.421	.599	.340	1

Note. The table reports weighted statistics for the positively worded (Pos1–Pos5) and negatively worded (Neg1–Neg5) items that had a five-category Likert response scale (1 = *does not apply at all*, 2 = *does rather not apply*, 3 = *partly*, 4 = *does rather apply*, 5 = *applies completely*). The correlations are based on pairwise complete observations. The overall sample size amounted to $n = 4,989$ students.

Model Results

The model results of the self-esteem scale in NEPS grade 5 are depicted in Table 3 to illustrate the main findings. The results of the other four analyses are qualitatively the same

and depicted in Appendices 3.1 to 3.4. As expected, the loadings for positively and negatively worded items showed opposite signs in the one-dimensional CFA. In the two-dimensional CFA, the loadings on the global factor showed the expected reverse-sign loading patterns across the mixed-worded items. The loadings of the negatively worded items on the method factor were of opposite sign to the corresponding loadings on the global factor.

Table 3

Model Estimation Results of the One-Dimensional (1-dim) and Two-Dimensional (2-dim) Confirmatory Factor Analyses (CFA) and Constrained Factor Mixture Analysis (FMA) for the Self-Esteem Scale in NEPS Grade 5 in Germany

	1-dim CFA		2-dim CFA				FMA			
	Global factor		Global factor		Method factor		Consistent class 86%		Inconsistent class 14%	
	Par.	SE	Par.	SE	Par.	SE	Par.	SE	Par.	SE
<i>Factors</i>										
Mean α	0		0		0		0		-0.81	0.56
Variance ϕ	1		1		1		1		1.67	0.20
<i>Loadings λ</i>										
Pos1	0.44	0.03	0.57	0.02			0.52	0.05	0.52	0.05
Pos2	0.37	0.03	0.56	0.02			0.47	0.03	0.47	0.03
Pos3	0.40	0.03	0.56	0.02			0.49	0.04	0.49	0.04
Pos4	0.47	0.04	0.62	0.03			0.54	0.04	0.54	0.04
Pos5	0.54	0.04	0.70	0.02			0.60	0.06	0.60	0.06
Neg1	-0.78	0.03	-0.41	0.03	0.71	0.03	-0.62	0.10	0.62	0.10
Neg2	-0.72	0.03	-0.40	0.04	0.63	0.05	-0.57	0.07	0.57	0.07
Neg3	-0.88	0.03	-0.52	0.03	0.77	0.04	-0.63	0.20	0.63	0.20
Neg4	-0.54	0.03	-0.18	0.04	0.63	0.04	-0.46	0.03	0.46	0.03
Neg5	-0.82	0.03	-0.49	0.03	0.71	0.04	-0.64	0.16	0.64	0.16
<i>Intercepts ν</i>										
Pos1	4.28	0.02	4.29	0.02			4.34	0.05	4.34	0.05
Pos2	4.08	0.02	4.08	0.02			4.13	0.05	4.13	0.05
Pos3	4.01	0.02	4.01	0.02			4.06	0.05	4.06	0.05
Pos4	3.89	0.03	3.89	0.03			3.95	0.06	3.95	0.06
Pos5	4.04	0.02	4.03	0.02			4.10	0.06	4.10	0.06
Neg1	2.45	0.03	2.45	0.03			2.30	0.07	3.92	0.44
Neg2	2.19	0.04	2.19	0.04			2.03	0.07	3.66	0.35
Neg3	1.84	0.03	1.84	0.03			1.55	0.11	4.21	0.39
Neg4	2.74	0.03	2.74	0.03			2.62	0.04	3.85	0.25
Neg5	1.80	0.02	1.80	0.02			1.62	0.10	3.44	0.55
<i>Res. var. Ψ</i>										
Pos1	0.55	0.03	0.42	0.02			0.43	0.02	0.43	0.02
Pos2	0.68	0.02	0.51	0.02			0.56	0.02	0.56	0.02
Pos3	0.77	0.03	0.62	0.02			0.65	0.02	0.65	0.02

Pos4	1.32	0.05	1.16	0.05	1.20	0.05	1.20	0.05
Pos5	0.83	0.04	0.64	0.03	0.70	0.04	0.70	0.04
Neg1	1.12	0.04	1.04	0.03	1.15	0.04	1.15	0.04
Neg2	1.20	0.05	1.16	0.05	1.20	0.05	1.20	0.05
Neg3	0.61	0.04	0.52	0.03	0.40	0.03	0.40	0.03
Neg4	1.58	0.04	1.44	0.04	1.56	0.05	1.56	0.05
Neg5	0.65	0.04	0.59	0.03	0.67	0.06	0.67	0.06

Note. The positively worded (Pos1–Pos5) and negatively worded (Neg1–Neg5) indicators had a five-category Likert response scale (1 = *does not apply at all*, 2 = *does rather not apply*, 3 = *partly*, 4 = *does rather apply*, 5 = *applies completely*). The table reports unstandardized parameter estimates and standard errors of model estimations in which sampling weights were applied. Fixed latent means and variances are depicted without standard errors. For the two-dimensional CFA model, the parameters for the orthogonal method factor are displayed in a separate column. For the constrained FMA model, the parameters are displayed for the classes of consistent versus inconsistent respondents separately. The reported class sizes are the final class proportions based on the estimated model in percent. Cases with missing values on all variables were excluded from the analyses. The effective sample size amounted to $n = 4,938$ students.

In line with our expectations, the final proportions of the inconsistent respondents based on the estimated models were much smaller than those of the consistent classes in all five datasets. The share of inconsistent respondents was especially small in the older samples (7% and 8% using the NEPS grade 9 sample or the PISA sample of fifteen-year-olds) compared to the younger samples (14% using the NEPS grade 5 sample and 12% and 20% using the PIRLS/TIMSS grade 4 sample). In the first dataset (self-esteem scale in NEPS grade 5, see Table 3), the intercepts of the negatively worded items were 2.02 on average and 4.12 for positively worded items in the consistent class. Hence, an individual with average self-esteem is expected to respond about one point either under or over the midpoint of the five-step scale depending on the negative or positive wording of the item. In the inconsistent class,

in contrast, the average item intercept for negatively worded items was estimated to be 3.82, which was almost as high as the result for the positively worded items. In this class, an individual with average self-esteem is therefore expected to agree almost as much with negatively worded items as with positively worded items. This is in line with the inconsistency in responding we expected—individuals did not switch sides in the response scale. In the other four generalization datasets, we found the same pattern (see Appendices 3.1 to 3.4).

The classes exhibited one of two things: Either the factor means were not statistically different (NEPS grade 5) or the latent means of the global factors indicated a lower average self-esteem, mathematics self-concept, and reading self-concept in the inconsistent as compared to the consistent classes.⁴ This implies that the average responses in the inconsistent classes were slightly closer to the midpoints of the scales. Therefore, respondents in the inconsistent classes were not only characterized by their opposite responses to negatively worded items, but also by less extreme positive self-attitudes (self-esteem, mathematics and reading self-concept), at least in four out of five datasets. The variances of the latent factors were either not statistically different in the inconsistent and consistent classes or they were higher in the former than in the latter. This implies that the inconsistent respondents gave equally homogeneous or more homogeneous responses across items.

We translated the FMA model estimates (see Table 3 and Appendices 3.1 to 3.4) into model-implied estimated average means, standard deviations, and item intercorrelations for the two types of items (see Table 4 and Appendices 4.1 to 4.4) by applying Wright's (1934) rules. As the model parameters (factor loadings, residual variances, and factor variances) in Table 3 and the average item intercorrelations in Table 4 show, the consistent class changed the side of the response scale they checked depending on the item wording in the NEPS grade

⁴ In Appendices 3.2 and 3.3, note that, in contrast to the NEPS and PISA scales, the PIRLS/TIMSS response scales went from agreement to disagreement. A higher value on the latent factor therefore reflects low self-concepts.

5 dataset. The constrained model forced the item intercorrelations to show positive intercorrelations between items with the same wording and negative intercorrelations between mixed-worded items. The magnitudes of these intercorrelations were about equal in size between these item groups. This reflects the modeled intercorrelation pattern (see Figure 1) and expected answering behavior, namely of responding to positively and negatively worded items in opposite ways. In the inconsistent class, the respondents did not change the side of the response scale they checked but tended to agree or disagree with all items, regardless of the wording. The model forced all item intercorrelations to be positive. The intercorrelations between items with the same wording and mixed-worded items were, however, estimated to be about equal in size. This also conforms to the expected intercorrelation pattern (see Figure 1) and is consistent with the answering behavior we expected, which entailed responding to positively and negatively worded items in the same way.

As the item intercepts in combination with the factor means in Table 3 and the average item means in Table 4 show, the consistent respondents scored about one scale point above the midpoint of the five-step response scale on the positively worded items and about one scale point below the scale's midpoint on the negatively worded items in NEPS grade 5. Therefore, the consistent respondents on average tended to agree with the positively worded items and to disagree with the negatively worded items and therefore expressed a rather positive self-esteem. The inconsistent respondents, in contrast, scored about at or slightly above the midpoint of the response scale on both the positively and negatively worded items. In this class, individuals are therefore expected to agree or disagree about as much with negatively worded items as with positively worded items. The items' standard deviations were slightly higher in the inconsistent than in the consistent class. We found qualitatively the same patterns of findings in the other datasets.

Table 4

Model-Implied Average Estimated Means, Standard Deviations, and Item Intercorrelations across Positively and Negatively Worded Items in the Latent Classes of the Constrained Factor Mixture Analysis (FMA) for the Self-Esteem Scale in NEPS Grade 5 in Germany

	Consistent class 86%		Inconsistent class 14%	
	Average Pos1–Pos5	Average Neg1–Neg5	Average Pos1–Pos5	Average Neg1–Neg5
<i>M</i>	4.12	2.02	3.69	3.34
<i>SD</i>	0.98	1.15	1.07	1.25
Item intercorrelation				
Average Pos1–Pos5	.289		.403	
Average Neg1–Neg5	-.283	.272	.392	.376

Note. The positively worded (Pos1–Pos5) and negatively worded (Neg1–Neg5) indicators had a five-category Likert response scale (1 = *does not apply at all*, 2 = *does rather not apply*, 3 = *partly*, 4 = *does rather apply*, 5 = *applies completely*).

Model fit. Table 5 depicts the model fit results of the one- and two-dimensional CFAs and the constrained FMA. As previous research on the dimensionality of mixed-worded scales has also found, the model fit statistics indicated a poor fit of the one-dimensional CFAs with the data in all five generalization analyses and a better fit of the two-dimensional CFAs. This implies that the method factors for the negatively worded items accounted for the typical item intercorrelation patterns associated with the mixed wording. Although they only had three parameters more than the two-dimensional CFA models, the constrained FMAs outperformed the two CFA models on both the AIC and BIC in all analyses. We interpret the better relative fit of the constrained FMA models as an indication that they fit the data at least as well as the traditional two-dimensional CFA solution in an absolute sense. The entropy values of the constrained FMAs ranged between .810 and .954 across the datasets, which we interpret as an indication for appropriate classification (Celeux & Soromenho, 1996). Note that the two-dimensional CFA model is not a variable-centered equivalent of the constrained FMA model. In Appendix 5, we describe the results of two less constrained FMA models. The main

conclusion of these additional models is that the rather strict parameter constraints in our main FMA model are necessary to identify the theoretical classes of interest in all five datasets. Given that, for our rather confirmatory study objectives, our latent classes are theoretically motivated and strictly constrained to reflect their definition in terms of consistent and inconsistent responding, we did not see it as useful or fruitful to opt for a more exploratory model specification.

Table 5

Model Fit Results of the One-Dimensional (1-dim) and Two-Dimensional (2-dim)

Confirmatory Factor Analyses (CFA), and the Constrained Factor Mixture Analysis

(Constrained FMA)

	1-dim CFA	2-dim CFA	Constrained FMA
<i>Self-esteem NEPS grade 5</i>			
<i>n</i>	4,938	4,938	4,938
No. par.	30	35	38
AIC	138,233	135,821	135,119
BIC	138,428	136,048	135,367
Entropy			.831
RMSEA	.100	.034	
CFI	.711	.971	
<i>Self-esteem NEPS grade 9</i>			
<i>n</i>	4,754	4,754	4,754
No. par.	30	35	38
AIC	117,870	116,014	115,873
BIC	118,064	116,241	116,118
Entropy			.882
RMSEA	.060	.039	
CFI	.857	.947	
<i>Math. self-concept PIRLS/TIMSS</i>			
<i>N</i>	5,885	5,885	5,885
No. par.	21	24	27
AIC	99,387	97,371	96,490
BIC	99,527	97,532	96,670
Entropy			.810
RMSEA	.099	.022	
CFI	.866	.995	
<i>Read. self-concept PIRLS/TIMSS</i>			
<i>n</i>	5,863	5,863	5,863
No. par.	21	24	27
AIC	98,249	97,251	94,364
BIC	98,389	97,411	94,545
Entropy			.954
RMSEA	.081	.042	
CFI	.841	.966	
<i>Read. self-concept PISA</i>			

<i>n</i>	4,703	4,703	4,703
No. par.	18	21	24
AIC	56,248	54,630	54,272
BIC	56,364	54,766	54,427
Entropy			.814
RMSEA	.168	.066	
CFI	.811	.981	

Validation of Latent Classes

To further substantiate the class definitions of the constrained FMA, we related the class membership to criterion variables. We argued that the person characteristics of reading abilities, cognitive skills, and carefulness would be positively associated with consistent responses to mixed-worded scales. The validity results are summarized in Table 6 for all datasets. We further explored overlaps in class membership across time and scales, as well as item response patterns within the consistent and inconsistent classes.

Reading achievement. We found a significant association between class membership and reading achievement in all five datasets. In the NEPS grade 5 sample, for example, students with a very low reading score ($z = -2$) were estimated to have a 42% probability of being in the inconsistent class compared to a probability of merely 9% and 1% for students with an average reading score ($z = 0$) or a very high reading score ($z = 2$). This association conformed to expectations.

Basic cognitive skills. In the NEPS grade 5 dataset—the only dataset including basic cognitive skills—cognitive reasoning was significantly related to class membership. This finding of lower cognitive skill scores in the inconsistent class conformed to expectations. We also found a negative association between the second basic cognitive skill variable—perceptual speed—and membership of the inconsistent class, but this was not significant. This did not conform to expectations.

Conscientiousness. In the NEPS grade 9 dataset, the only dataset including conscientiousness scores, we did not find a significant association between class membership

and conscientiousness, a variable that we used as a proxy for carefulness. The insignificant finding therefore did not conform to our expectations.

Table 6

Descriptive Statistics of the Validity Covariates and Results of the Validity Analyses Regarding the Association of the Inconsistent Class Membership and Reading, Cognitive Skills (Reasoning and Perceptual Speed), and Conscientiousness in all Available Datasets

	Descriptive statistics of predictor variables			Predictor of class membership in FMA						
	<i>M</i>	<i>SD</i>	Miss- ing values	<i>n</i>	<i>b</i>	<i>SE</i>	<i>p</i>	Probability inconsistent class member by <i>z</i> -score		
								<i>z</i> (-2)	<i>z</i> (0)	<i>z</i> (2)
<i>Reading</i>										
Self-est. NEPS gr. 5	-0.06	1.25	0%	4,924	-0.96	0.13	< .001	42%	9%	1%
Self-est. NEPS gr. 9	-0.14	1.12	11%	4,229	-1.01	0.25	< .001	28%	5%	1%
Math. self-con. PIRLS/TIMSS	510.40	89.19	0%	5,885	-1.05	0.07	< .001	43%	9%	1%
Read. self-con. PIRLS/TIMSS	510.40	89.19	0%	5,863	-1.27	0.09	< .001	70%	16%	1%
Read. self-con. PISA	505.41	108.31	0%	4,703	-1.75	0.10	< .001	62%	5%	0%
<i>Reasoning</i>										
Self-est. NEPS gr. 5	6.82	2.60	0%	4,919	-0.43	0.10	< .001	26%	13%	6%
<i>Perc. speed</i>										
Self-est. NEPS gr. 5	43.97	13.45	0%	4,934	-0.15	0.10	.129	17%	14%	10%
<i>Conscientious.</i>										
Self-est. NEPS gr. 9	3.04	0.85	2%	4,680	0.00	0.13	.996	7%	7%	7%

Note. Cases with missing values on all variables were excluded from the analyses. The table reports the effective sample sizes, the unstandardized descriptive statistics of the validity covariates, and the unstandardized parameter and standard error estimates from the validity analyses, in which the validity covariates were *z*-standardized and predicted the latent class variable in the constrained factor mixture analysis (FMA) models. The table also depicts the probability of being in the inconsistent class for three *z*-values of the predictor variables.

Overlap in class membership. In an initial exploration, we investigated the overlap between class membership and the inconsistent class across time or across scales within the

same questionnaire. This type of tentative exploration allowed us to gain some insights into the temporal stability of inconsistent responses, which we assumed would be related to person characteristics. In the NEPS study, some fifth-graders were followed up in grade 9 and then filled in the same mixed-worded self-esteem scale. The inconsistent classes consisted of 14% of respondents in grade 5 and 7% of those in grade 9. Of the individuals who were identified as inconsistent respondents in grade 5, 8% were also in the inconsistent class in grade 9. Hence, this probability of being in the inconsistent classes across grades was almost the same as the 7% baseline membership probability for grade 9. Therefore, this finding did not conform to our expectations.

In the joint PIRLS/TIMSS survey, students responded first to a mixed-worded reading scale and then to a mathematics self-concept scale. Of the students who were classified as belonging to the inconsistent class with regard to the reading self-concept scale, 24% were also identified as inconsistent respondents to the mathematics self-concept scale. This was twice as high as the 12% baseline probability of belonging to the inconsistent class with regard to the mathematics self-concept scale across all respondents. This finding was consistent with expectations, and although the probability to belong to the inconsistent class was still low, we interpret it cautiously as an indication of a (more than random) overlap between inconsistent respondents across scales in the same survey.

Response patterns. We validated the classification of consistent versus inconsistent respondents in a simple response pattern analysis for the working example of NEPS grade 5. We extracted the most likely class membership from the constrained FMA results for all respondents with complete responses to all 10 items ($n = 4,254$) and identified those with prototypical responses. We flagged consistent respondents as those who agreed (category 4 or 5) with all positively and disagreed (category 1 or 2) with all negatively worded items, or, vice versa, disagreed (category 1 or 2) with all positively and agreed (category 4 or 5) with all negatively worded items. Conform to expectations, all of these $n = 618$ prototypically

consistent respondents belonged to the consistent class in the FMA. We also flagged all prototypically inconsistent respondents who either agreed (category 4 or 5) or disagreed (category 1 or 2) with all ten items and, conform to expectations, found these $n = 43$ to be assigned to the inconsistent class in the FMA. The $n = 15$ respondents who chose the middle category (category 3) for all ten items were identified as consistent respondents in the FMA. Even though middle responses imply that the respondents do not change the side of the response scale, this does not qualify as inconsistent response from a theoretical perspective because they report a medium self-esteem on all items. Therefore, this finding conforms to expectations, as well, so that all response pattern analyses anecdotally support the FMA classification.

Overall, we found 3,012 different response patterns of those that were classified as consistent respondents ($n = 3,794$) and 439 different response patterns of those classified as inconsistent respondents in the NEPS grade 5 FMA ($n = 460$). In the consistent class, the three most common response patterns were 5555511111 ($n = 172$), 5455511111 ($n = 30$), and 5555511121 ($n = 25$). In the inconsistent class, the three most common response patterns were 5555555555 ($n = 14$), 4444444444 ($n = 4$), and 4544444344 ($n = 2$).

Discussion

The aim of the present study was to identify two groups of respondents—consistent versus inconsistent respondents—to mixed-worded scales using a constrained FMA model. In five datasets, the constrained FMA models identified larger consistent classes and smaller inconsistent classes. The estimated item means and item intercorrelations in the classes conformed to expectations. In other words, the consistent classes contained respondents who changed sides according to the item wording, and the inconsistent classes captured respondents who continued checking the same side. The shares of inconsistent respondents ranged between 12% and 20% in the younger samples of fourth- and fifth-graders. In the older samples of ninth-graders and fifteen-year-olds, the shares of inconsistent respondents

were 7% and 8%. Inconsistent respondents were therefore more prevalent among younger age groups than older ones in our five analyses.

Although we constrained our FMA model explicitly to identify consistent and inconsistent respondents and the model results conformed to expectations, the risk of identifying spurious classes in mixture models remains (cf. Bauer & Curran, 2004). To minimize this risk, we ran a number of validity analyses. In line with our hypotheses and previous research (Gnambs & Schroeders, 2017; Marsh, 1986; Melnick & Gable, 1990; Steedle et al., 2019), we found that the probability of belonging to the inconsistent response class was much higher for poor readers and also for students with low cognitive reasoning scores. There was an association between membership of the inconsistent class and lower processing speed, a second cognitive basic skill variable, but this association was not significant. Additionally, our expectation that the inconsistent response behavior would be related to low conscientiousness scores was not substantiated. The conscientiousness scale in the NEPS study was, however, a mean score of one positively and one negatively worded item, after simply reversing the negatively worded item. Therefore, the conscientiousness scale might be biased by inconsistent responding towards mixed-worded items, in the same way as occurs in the self-esteem scale. Furthermore, this two-item scale was possibly unreliable. This might be explanations for why we did not find a significant association between class membership and the conscientiousness scale. Since the inconsistent classes showed lower average self-attitudes (self-esteem, mathematics and reading self-concept) than the consistent classes in four datasets, as indicated by the factor means, the associations between class membership and the criterion variables could, however, be confounded.

Contrary to our expectations, inconsistent respondents in grade 5 were not more likely to belong to the inconsistent class in grade 9 than all respondents, independent of the previous class membership. Arguably, a pronounced overlap could have been expected if the interval had been shorter than four years. Yet, a previous study found pronounced stability for an

item-wording-related method factor over four years between grades 8 and 12 (Motl & DiStefano, 2002). More in line with expectations, we found that the probability of belonging to the inconsistent class was twice as high for those who were also identified as inconsistent respondents to another mixed-worded scale in the same questionnaire. This finding has to be interpreted with caution, however, since the overlap in class memberships was not very pronounced and only investigated in one dataset.

We furthermore compared the class membership in the NEPS grade 5 FMA for individuals with prototypically consistent or inconsistent response patterns. In contrast to the FMA model, this additional flagging method used simple recoding of directly observed, prototypical response patterns. Conform to our expectations, all respondents who gave prototypically consistent responses were also classified as consistent respondents, and all who gave prototypically inconsistent responses were classified as inconsistent respondents in the constrained FMA.

Limitations

Assuming that our constrained FMA actually identified consistent and inconsistent respondents, we still see important limitations of the interpretability of our findings. Our analyses did not allow us to disentangle the two aforementioned reasons for inconsistent response behavior, a lack of skills versus carelessness. For instance, careless inconsistent respondents might also respond inattentively to the achievement tests and therefore score low on the reading and reasoning variables. Future research could investigate the roots of inconsistent responding in more detail. It would, for example, be interesting to see if the inconsistent respondents identified in the constrained FMA approach would also be flagged as careless respondents when using other methods like instructed-response items or person-fit statistics (see e.g., Hong et al., 2019; Kam & Chan, 2018; Meade & Craig, 2012; Patton et al., 2019).

Furthermore, it is important to stress that there are alternative explanations for the phenomenon under investigation. Another explanation for the item intercorrelation irregularities in mixed-worded scales is that positively and negatively worded items indeed measure substantively different constructs. Owens (1994), for example, concluded that the positively worded items of Rosenberg's Self-Esteem Scale measure "positive self-worth," while the negatively worded items measure "self-deprecation" (p. 396). This view has been challenged in the past because the phenomenon applies to various mixed-worded scales and the constructs showed no differential correlations with external variables (cf. e.g., Greenberger et al., 2003; Marsh, 1996). We therefore followed an argument based on wording-related method effects. Our study did not investigate other item wording effects that might be in place and that might be captured by the method factor in the two-dimensional CFA or by other types of mixture models. The term *method effects*, as we use it here, summarizes all sources of systematic variance that are independent of the substantive constructs of interest (e.g., Podsakoff et al., 2003).

Limitations to the generalizability of the findings. Although we replicated our findings across a range of mixed-worded scales and populations, there are still important limitations of the generalizability of the findings. First, our study focused on four mixed-worded attitudinal scales, each of which comprised six to ten items with fixed Likert response scales and aimed to measure one underlying construct. It would be interesting to investigate scales on constructs other than attitudes, scales with even more items, or scales that aimed at measuring multiple constructs in future research. Previous studies that used other types of mixed-worded scales also exhibited item wording effects. These include a study on a mixed-worded job diagnostic survey that was designed to have five dimensions (Idaszak & Drasgow, 1987; cf. also Cordery & Sevastos, 1993; Harvey et al., 1985). In our study, we found that reading achievement was a stronger predictor for class membership in the two datasets with reading self-concept scales than in the others (see Table 6). Future research could more

systematically investigate the association of reading and mixed-worded scales that are more or less distant to reading achievement. Furthermore, our limited number of scales meant we could not investigate the effects of further characteristics of mixed-worded scales, like the order of positively and negatively worded items, the extremity of item wording, the effect of reversing the response scale or of using other response scale formats (see e.g., Barnette, 2000; McPherson & Mohr, 2005). Additionally, the negatively worded items in this study were not parallel versions of the positively worded items. All these item and scale characteristics could be systematically manipulated and investigated in more detail in future experimental studies (see e.g., Barnette, 2000; Benson & Hocesvar, 1985; Greenberger et al., 2003).

Second, we studied samples from three countries and two age groups, since previous research suggested that the item wording effect might be more or less pronounced in different countries and age groups (e.g., Lindwall et al., 2012; Marsh, 1986; Marsh et al., 2013; Weems et al., 2003). Future research could, however, investigate the prevalence of inconsistent respondents in different age and country samples more systematically. If the inconsistent response behavior is indeed rooted in difficulties in understanding the item and adjusting the answers accordingly, the degree of difficulty could vary between children, adolescents, and adults or between speakers of languages that handle negative wording differently.

Third, we only investigated mixed-worded attitude scales that were assessed in low-stakes large-scale assessments. If the inconsistent response behavior indeed related to careless responding, it might occur less often in high-stakes situations like job assessment questionnaires, for example. This could be another valuable area for future research.

Methodological limitations. The present study made a number of strong assumptions regarding two latent classes of consistent and inconsistent respondents to unidimensional mixed-worded scales. We assumed that the consistent class would respond consistently and the inconsistent class would respond inconsistently to the wording across all items. Further, we modeled both types of respondents as showing the same responses to the positively

worded items (same intercepts and factor loadings), that is, the scalar invariance of the positively worded items across latent classes. We modeled the inconsistent respondents as giving opposite responses to the negatively worded items in comparison to the consistent respondents (factor loadings with equal value and opposite sign), that is, metric equivalence with reversed signs. We modeled these strict assumptions of prototypically consistent and inconsistent responses in order to characterize the constrained FMA model and illustrate our approach. It might be worth exploring further options for relaxing these model constraints. This could, however, impair the interpretability of the findings (cf., Appendix 5) or risk empirical underidentification and corresponding convergence issues (cf. footnote 1). We would further like to acknowledge that our study was limited to investigating a specific kind of population heterogeneity, namely consistent and inconsistent respondents. We did not explore further population heterogeneity that might exist.

Conclusions for the Use of Mixed-Worded Scales

The present study suggests two main conclusions with regard to using mixed-worded questionnaire scales. On the one hand, implementing the mixed item wording allows researchers to identify consistent and inconsistent respondents to positively and negatively worded items (e.g., Hong et al., 2019; Patton et al., 2019; Steedle et al., 2019). This allows them to investigate the characteristics of the inconsistent respondents, as in the present study, or to run analyses on a cleaned dataset that solely contains consistent respondents. This may be a relevant measure to control the quality of a dataset.

On the other hand, using mixed item wording to improve the psychometric properties of questionnaire scales (see e.g., Benson & Hocevar, 1985; Likert, 1974; Nunnally & Bernstein, 1994) is, according to our study, problematic. The findings here, along with previous findings, suggest that the mixed wording introduces method effects and therefore systematic measurement error (see e.g., Barnette, 2000; Carmines & Zeller, 1974; Marsh, 1986, 1996; Podsakoff et al., 2003). Mixed-worded scales seem thus not suitable for

measuring unidimensional latent constructs, at least under the given circumstances (i.e., low stakes assessments, young respondents, etc.) and if the method effects are not accounted for (i.e., as in our constrained factor mixture model, for instance). When not considering method effects, a simple mean scale score that averages the responses after reversing the scores of the negatively worded items would be systematically biased for inconsistent respondents, since mean scale scores require that the positively and negatively worded items work in diametrically opposite ways. If inconsistent respondents actually tended to have lower reading or cognitive skills scores, as suggested by our study, this systematic bias would furthermore relate to important respondent characteristics. This again would bias relationships between the mean scores and other variables in applied research. Furthermore, our study indicated that the inconsistent responding was rather stable across mixed-worded scales in the same questionnaire. Therefore, different mixed-worded scales would correlate more due to common item wording effects.

Nevertheless, simple mean scale scores of mixed-worded scales are still in common use. They are, for example, provided in the NEPS data files that we used in the present study and that are prepared for secondary analyses. When investigating applied research questions—that is, questions that do not concern consistent versus inconsistent respondents—we clearly recommend not using these biased mean scale scores. When conducting secondary analyses on the basis of large-scale datasets, we would recommend computing mean scale scores for the positively worded items alone (see also Benson & Hocevar, 1985; Marsh, 1986). When developing new questionnaire scales, we would suggest setting up scales with only one type of item wording, especially for samples with young respondents and those with low reading abilities (see e.g., Dunbar et al., 2000; Lindwall et al., 2012; Marsh, 1996; Melnick & Gable, 1990; Weems et al., 2003). While solely using positively worded items may circumvent the issue of mixed item wording effects in the data, it does not, however, avoid the problem of careless responding.

Conclusions for the Use of the Constrained Factor Mixture Model

Our findings have a number of implications for using constrained FMA models to identify consistent and inconsistent respondents. First, the identification of consistent and inconsistent respondents requires that the mixed-worded scales have a clear and strong factor structure. It is easier to separate latent classes based on different patterns of factor loadings if these loadings are pronounced (negative or positive). If even the items with the same wording do not intercorrelate highly, the inconsistent response pattern cannot be meaningfully distinguished from responding to independent items. A second important issue is the number of items. In our five datasets, we investigated mixed-worded attitudinal scales with between 6 and 10 items and 4 and 5 Likert response categories. We believe that applying our approach would be difficult in scales with even fewer items and fewer response categories. Third, we used datasets with more than 4,500 respondents to the mixed-worded scales. Since many parameters have to be estimated and the inconsistent classes can be quite small, applying this approach to small datasets will probably be difficult and lead to unreliable estimates.

The major advantage of our FMA approach as compared to traditional CFAs is that we model very specific expected response behavior characteristics of two types of respondents. Such confirmatory approaches have to be well-supported by theoretical assumptions, which should guide all constraining decisions. We recommend to run alternative more and less constrained models to scrutinize the comparative findings. FMA models are complex and a deep understanding of each model parameter is critical to develop useful and parsimonious models. Although confirmatory models do not necessarily aim at attaining the best possible model fit, evaluating the fit to the data is important. The models should fit at least as well as substantive competitor models (like the two-dimensional CFAs in this study). Furthermore, the entropy should be high enough to suggest an adequate classification accuracy (cf. e.g., Celeux & Soromenho, 1996). Another important aspect is that the validity of the classification should be scrutinized (like with relevant criterion variable approaches, overlap, and response

pattern analyses in this study). Because of the risk of identifying spurious latent classes, we strongly discourage the stand-alone use of constrained FMAs without comparisons to competitor models, model fit evaluations, and validity checks.

Central Assets of the Present Study

Our study connected two strands of previously unrelated research. In one strand, studies have investigated the dimensionality of mixed-worded scales and found item wording effects by applying CFA approaches (e.g., DiStefano & Motl, 2006, 2009; Marsh et al., 2010; Yang et al., 2012). These did not make it possible to investigate the phenomenon underlying wording-related method effects in a straightforward way. Furthermore, they assumed population homogeneity, i.e., that the item wording effect would apply to all respondents to a certain degree. Findings showing that the mixed item wording effect phenomenon relates to person characteristics are therefore very difficult to interpret. Another strand of research has used mixed-worded scales as a tool to identify consistent and inconsistent respondents, without modeling the substantive latent constructs (Kam & Chan, 2018; Melnick & Gable, 1990; Steedle et al., 2019). These studies illustrated different characteristics of consistent and inconsistent respondents.

The present study made use of the capacities of mixture modeling. Mixture approaches are increasingly showing their potential to investigate population heterogeneity issues in response behavior while modeling the underlying substantive constructs at the same time (see e.g., Jin et al., 2018; Kam, 2018; Kam & Fan, 2018; Meade & Craig, 2012; Mislevy & Verhelst, 1990; Sen & Cohen, 2019; Tijmstra et al., 2018). We applied a constrained mixture model to our research problem that allowed us to investigate both the substantive attitude constructs and consistent versus inconsistent respondents towards mixed-worded scales at the same time. Our study assumed a clear mechanism: Some respondents answered negatively worded items as if they were positively worded, while most respondents answered positively and negatively worded items in opposite ways. The latent classes of consistent and

inconsistent respondents showed many expected properties, also in terms of differential respondent characteristics.

References

- Asparouhov, T., & Muthén, B. (2014). *Auxiliary variables in mixture modeling: 3-step approaches using Mplus*. Mplus Web Notes No. 15.
<https://www.statmodel.com/download/webnotes/webnote15.pdf>
- Barnette, J. J. (2000). Effects of stem and Likert response option reversals on survey internal consistency: If you feel the need, there is a better alternative to using those negatively worded stems. *Educational and Psychological Measurement*, 60(3), 361–370.
<https://doi.org/10.1177/00131640021970592>
- Bauer, D. J., & Curran, P. J. (2004). The integration of continuous and discrete latent variable models: Potential problems and promising opportunities. *Psychological Methods*, 9(1), 3–29. <https://doi.org/10.1037/1082-989X.9.1.3>
- Benson, J., & Hocevar, D. (1985). The impact of item phrasing on the validity of attitude scales for elementary school children. *Journal of Educational Measurement*, 22(3), 231–240. <https://doi.org/10.1111/j.1745-3984.1985.tb01061.x>
- Blossfeld, H.-P., Roßbach, H.-G., & von Maurice, J. (Eds.). (2011). *Education as a lifelong process: The German national educational panel study (NEPS)*.
- Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association*, 71(356), 791–799. <https://doi.org/10.1080/01621459.1976.10480949>
- Carmines, E. G., & Zeller, R. A. (1974). On establishing the empirical dimensionality of theoretical terms: An analytical example. *Political Methodology*, 1(4), 75–96.
- Celeux, G., & Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification*, 13(2), 195–212.
<https://doi.org/10.1007/BF01246098>
- Cole, D. A., Ciesla, J. A., & Steiger, J. H. (2007). The insidious effects of failing to include design-driven correlated residuals in latent-variable covariance structure analysis. *Psychological Methods*, 12(4), 381–398. <https://doi.org/10.1037/1082-989X.12.4.381>
- Cordery, J. L., & Sevastos, P. P. (1993). Responses to the original and revised Job Diagnostic Survey: Is education a factor in responses to negatively worded items? *The Journal of Applied Psychology*, 78(1), 141–143. <https://doi.org/10.1037/0021-9010.78.1.141>
- Courvoisier, D. S., Eid, M., & Nussbeck, F. W. (2007). Mixture distribution latent state-trait analysis: Basic ideas and applications. *Psychological Methods*, 12(1), 80–104.
<https://doi.org/10.1037/1082-989X.12.1.80>

- Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods*, 1(1), 16–29. <https://doi.org/10.1037/1082-989X.1.1.16>
- DiStefano, C., & Motl, R. W. (2006). Further investigating method effects associated with negatively worded items on self-report surveys. *Structural Equation Modeling: A Multidisciplinary Journal*, 13(3), 440–464. https://doi.org/10.1207/s15328007sem1303_6
- DiStefano, C., & Motl, R. W. (2009). Self-esteem and method effects associated with negatively worded items: Investigating factorial invariance by sex. *Structural Equation Modeling: A Multidisciplinary Journal*, 16(1), 134–146. <https://doi.org/10.1080/10705510802565403>
- Dunbar, M., Ford, G., Hunt, K., & Der, G. (2000). Question wording effects in the assessment of global self-esteem. *European Journal of Psychological Assessment*, 16(1), 13–19. <https://doi.org/10.1027//1015-5759.16.1.13>
- Foy, P. (2013). *TIMSS and PIRLS 2011 user guide for the fourth grade combined international database*. International Study Center; Lynch School of Education; Boston College.
- Gehrer, K., Zimmermann, S., Artelt, C., & Weinert, S. (2012). *The assessment of reading competence (including sample items for grade 5 and 9): Scientific use file 2012, version 1.0.0*. University of Bamberg, National Educational Panel Study.
- Gnambs, T., & Schroeders, U. (2017). Cognitive abilities explain wording effects in the Rosenberg Self-Esteem Scale. *Assessment*, 1–15. <https://doi.org/10.1177/1073191117746503>
- Greenberger, E., Chen, C., Dmitrieva, J., & Farruggia, S. P. (2003). Item-wording and the dimensionality of the Rosenberg Self-Esteem Scale: Do they matter? *Personality and Individual Differences*, 35(6), 1241–1254. [https://doi.org/10.1016/S0191-8869\(02\)00331-8](https://doi.org/10.1016/S0191-8869(02)00331-8)
- Haberkorn, K., & Pohl, S. (2013). *Cognitive basic skills – Data in the scientific use file*. University of Bamberg, National Educational Panel Study.
- Hallquist, M., Wiley, J., & van Lissa, C. (2018). *MplusAutomation: An R package for facilitating large-scale latent variable analyses in Mplus*. R package version 0.7-3. <https://github.com/michaelhallquist/MplusAutomation>
- Harvey, R. J., Billings, R. S., & Nilan, K. J. (1985). Confirmatory factor analysis of the Job Diagnostic Survey: Good news and bad news. *Journal of Applied Psychology*, 70(3), 461–468. <https://doi.org/10.1037/0021-9010.70.3.461>

- Hong, M., Steedle, J. T., & Cheng, Y. (2019). Methods of detecting insufficient effort responding: Comparisons and practical recommendations. *Educational and Psychological Measurement*, 1–34. <https://doi.org/10.1177/0013164419865316>
- Idaszak, J. R., & Drasgow, F. (1987). A revision of the Job Diagnostic Survey: Elimination of a measurement artifact. *Journal of Applied Psychology*, 72(1), 69–74. <https://doi.org/10.1037/0021-9010.72.1.69>
- Jin, K.-Y., Chen, H.-F., & Wang, W.-C. (2018). Mixture item response models for inattentive responding behavior. *Organizational Research Methods*, 21(1), 197–225. <https://doi.org/10.1177/1094428117725792>
- Kam, C. C. S. (2018). Testing the assumption of population homogeneity in the measurement of dispositional optimism: Factor mixture modeling analysis. *Journal of Personality Assessment*, 1–12. <https://doi.org/10.1080/00223891.2018.1502194>
- Kam, C. C. S., & Chan, G. H.-h. (2018). Examination of the validity of instructed response items in identifying careless respondents. *Personality and Individual Differences*, 129, 83–87. <https://doi.org/10.1016/j.paid.2018.03.022>
- Kam, C. C. S., & Fan, X. (2018). Investigating response heterogeneity in the context of positively and negatively worded items by using factor mixture modeling. *Organizational Research Methods*, 9, 1–20. <https://doi.org/10.1177/1094428118790371>
- Likert, R. (1974). The method of constructing an attitude scale. In G. M. Maranell (Ed.), *Methodological perspectives. Scaling: A sourcebook for behavioral scientists* (pp. 233–243). Aldine Publ.
- Lindwall, M., Barkoukis, V., Grano, C., Lucidi, F., Raudsepp, L., Liukkonen, J., & Thøgersen-Ntoumani, C. (2012). Method effects: The problem with negatively versus positively keyed items. *Journal of Personality Assessment*, 94(2), 196–204. <https://doi.org/10.1080/00223891.2011.645936>
- Lubke, G. H., & Luningham, J. (2017). Fitting latent variable mixture models. *Behaviour Research and Therapy*, 98, 91–102. <https://doi.org/10.1016/j.brat.2017.04.003>
- Lubke, G. H., & Muthén, B. (2005). Investigating population heterogeneity with factor mixture models. *Psychological Methods*, 10(1), 21–39. <https://doi.org/10.1037/1082-989X.10.1.21>
- Magazine, S. L., Williams, L. J., & Williams, M. L. (1996). A confirmatory factor analysis examination of reverse coding effects in Meyer and Allen's Affective and Continuance

- Commitment Scales. *Educational and Psychological Measurement*, 56(2), 241–250.
<https://doi.org/10.1177/0013164496056002005>
- Marsh, H. W. (1986). Negative item bias in ratings scales for preadolescent children: A cognitive-developmental phenomenon. *Developmental Psychology*, 22(1), 37–49.
<https://doi.org/10.1037/0012-1649.22.1.37>
- Marsh, H. W. (1996). Positive and negative global self-esteem: A substantively meaningful distinction or artifacts? *Journal of Personality and Social Psychology*, 70(4), 810–819.
<https://doi.org/10.1037/0022-3514.70.4.810>
- Marsh, H. W., Abduljabbar, A. S., Abu-Hilal, M. M., Morin, A. J. S., Abdelfattah, F., Leung, K. C., Xu, M. K., Nagengast, B., & Parker, P. (2013). Factorial, convergent, and discriminant validity of TIMSS math and science motivation measures: A comparison of Arab and Anglo-Saxon countries. *Journal of Educational Psychology*, 105(1), 108–128.
<https://doi.org/10.1037/a0029907>
- Marsh, H. W., Scalas, L. F., & Nagengast, B. (2010). Longitudinal tests of competing factor structures for the Rosenberg Self-Esteem Scale: Traits, ephemeral artifacts, and stable response styles. *Psychological Assessment*, 22(2), 366–381.
<https://doi.org/10.1037/a0019225>
- Masyn, K. E., Henderson, C. E., & Greenbaum, P. E. (2010). Exploring the latent structures of psychological constructs in social development using the dimensional-categorical spectrum. *Social Development*, 19(3), 470–493. <https://doi.org/10.1111/j.1467-9507.2009.00573.x>
- McPherson, J., & Mohr, P. (2005). The role of item extremity in the emergence of keying-related factors: An exploration with the life orientation test. *Psychological Methods*, 10(1), 120–131. <https://doi.org/10.1037/1082-989X.10.1.120>
- Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, 17(3), 437–455. <https://doi.org/10.1037/a0028085>
- Melnick, S. A., & Gable, R. K. (1990). The use of negative item stems: A cautionary note. *Educational Research Quarterly*, 14(3), 31–36.
- Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58(4), 525–543. <https://doi.org/10.1007/BF02294825>
- Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika*, 55(2), 195–215.
<https://doi.org/10.1007/BF02295283>

- Motl, R. W., & DiStefano, C. (2002). Longitudinal invariance of self-esteem and method effects associated with negatively worded items. *Structural Equation Modeling: A Multidisciplinary Journal*, 9(4), 562–578. https://doi.org/10.1207/S15328007SEM0904_6
- Mullis, I. V. S., Martin, M. O., Kennedy, A. M., Trong, K. L., & Sainsbury, M. (2009). *PIRLS 2011 assessment framework*. International Study Center, Boston College, Lynch School of Education.
- Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., & Preuschoff, C. (2009). *TIMSS 2011 assessment frameworks*. International Study Center, Boston College, Lynch School of Education.
- Muthén, B. O., & Satorra, A. (1995). Complex sample data in structural equation modeling. *Sociological Methodology*, 25, 267. <https://doi.org/10.2307/271070>
- Muthén, L. K., & Muthén, B. O. (1998–2017). *Mplus user's guide* (8th edition). Muthén & Muthén.
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd edition). McGraw-Hill series in psychology. McGraw-Hill.
<http://www.loc.gov/catdir/description/mh022/93022756.html>
- OECD. (2019). *PISA 2018 assessment and analytical framework*. OECD Publishing.
- Owens, T. J. (1994). Two dimensions of self-esteem: Reciprocal effects of positive self-worth and self-deprecation on adolescent problems. *American Sociological Review*, 59(3), 391–407.
- Patton, J. M., Cheng, Y., Hong, M., & Diao, Q. (2019). Detection and treatment of careless responses to improve item parameter estimation. *Journal of Educational and Behavioral Statistics*, 44(3), 309–341. <https://doi.org/10.3102/1076998618825116>
- Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *The Journal of Applied Psychology*, 88(5), 879–903.
<https://doi.org/10.1037/0021-9010.88.5.879>
- Quilty, L. C., Oakman, J. M., & Risko, E. (2006). Correlates of the Rosenberg Self-Esteem Scale method effects. *Structural Equation Modeling: A Multidisciplinary Journal*, 13(1), 99–117. https://doi.org/10.1207/s15328007sem1301_5
- Reise, S. P., Kim, D. S., Mansolf, M., & Widaman, K. F. (2016). Is the bifactor model a better model or is it just better at modeling implausible responses? Application of iteratively

- reweighted least squares to the Rosenberg Self-Esteem Scale. *Multivariate Behavioral Research*, 51(6), 818–838. <https://doi.org/10.1080/00273171.2016.1243461>
- Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton University Press.
- Schmitt, N., & Stults, D. M. (1985). Factors defined by negatively keyed items: The result of careless respondents? *Applied Psychological Measurement*, 9(4), 367–373. <https://doi.org/10.1177/014662168500900405>
- Sen, S., & Cohen, A. S. (2019). Applications of mixture IRT models: A literature review. *Measurement: Interdisciplinary Research and Perspectives*, 17(4), 177–191. <https://doi.org/10.1080/15366367.2019.1583506>
- Steedle, J. T., Hong, M., & Cheng, Y. (2019). The effects of inattentive responding on construct validity evidence when measuring social–emotional learning competencies. *Educational Measurement: Issues and Practice*, 38(2), 101–111. <https://doi.org/10.1111/emip.12256>
- Tijmstra, J., Bolsinova, M., & Jeon, M. (2018). General mixture item response models with different item response structures: Exposition with an application to Likert scales. *Behavior Research Methods*, 50(6), 2325–2344. <https://doi.org/10.3758/s13428-017-0997-0>
- Tomás, J. M., Oliver, A., Galiana, L., Sancho, P., & Lila, M. (2013). Explaining method effects associated with negatively worded items in trait and state global and domain-specific self-esteem scales. *Structural Equation Modeling: A Multidisciplinary Journal*, 20(2), 299–313. <https://doi.org/10.1080/10705511.2013.769394>
- Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis*, 18(4), 450–469. <https://doi.org/10.1093/pan/mpq025>
- Weems, G. H., Onwuegbuzie, A. J., Schreiber, J. B., & Eggers, S. J. (2003). Characteristics of respondents who respond differently to positively and negatively worded items on rating scales. *Assessment & Evaluation in Higher Education*, 28(6), 587–606. <https://doi.org/10.1080/0260293032000130234>
- Woods, C. M. (2006). Careless responding to reverse-worded items: Implications for confirmatory factor analysis. *Journal of Psychopathology and Behavioral Assessment*, 28(3), 186–191. <https://doi.org/10.1007/s10862-005-9004-7>
- Wright, S. (1934). The method of path coefficients. *The Annals of Mathematical Statistics*, 5(3), 161–215. <https://doi.org/10.1214/aoms/1177732676>

Yang, Y., Chen, Y.-H., Lo, W.-J., & Turner, J. E. (2012). Cross-cultural evaluation of item wording effects on an attitudinal scale. *Journal of Psychoeducational Assessment*, 30(5), 509–519. <https://doi.org/10.1177/0734282911435461>

Appendix

Appendix 1: Item Wording, Descriptive Statistics, and Item Intercorrelations of the Mixed-Worded Scales

Appendix 1.1

*German Item Wording, Descriptive Statistics, and Item Intercorrelations of the Self-Esteem Scale in NEPS**Grade 9 in Germany in Original Sequence (Top) and Ordered by Item Type (Bottom)*

Items	Inwieweit treffen folgende Aussagen auf dich zu?	<i>M</i>	<i>SD</i>	Skewness	Excess kurtosis	Missing values
Pos1	a) Alles in allem bin ich mit mir selbst zufrieden.	3.77	0.98	-0.62	0.05	1%
Neg1	b) Hin und wieder denke ich, dass ich gar nichts taue.	2.48	1.11	0.37	-0.66	1%
Pos2	c) Ich besitze eine Reihe guter Eigenschaften.	3.82	0.85	-0.45	-0.06	1%
Pos3	d) Ich kann vieles genauso gut wie die meisten anderen Menschen auch.	3.80	0.86	-0.55	0.25	1%
Neg2	e) Ich fürchte, es gibt nicht viel, worauf ich stolz sein kann.	2.26	1.09	0.71	-0.19	1%
Neg3	f) Ich fühle mich von Zeit zu Zeit richtig nutzlos.	2.12	1.15	0.84	-0.18	1%
Pos4	g) Ich halte mich für einen wertvollen Menschen, jedenfalls bin ich nicht weniger wertvoll als andere auch.	3.68	1.07	-0.61	-0.23	2%
Neg4	h) Ich wünschte, ich könnte vor mir selbst mehr Achtung haben.	2.61	1.14	0.19	-0.79	3%
Neg5	i) Alles in allem neige ich dazu, mich für eine Versagerin oder einen Versager zu halten.	1.99	1.11	0.97	0.09	2%
Pos5	j) Ich habe eine positive Einstellung zu mir selbst gefunden.	3.67	1.04	-0.50	-0.43	2%

Item intercorrelations										
	Pos1	Pos2	Pos3	Pos4	Pos5	Neg1	Neg2	Neg3	Neg4	Neg5
Pos1	1									
Pos2	.486	1								
Pos3	.453	.623	1							
Pos4	.432	.470	.437	1						
Pos5	.634	.501	.474	.520	1					
Neg1	-.459	-.332	-.328	-.357	-.486	1				
Neg2	-.432	-.378	-.358	-.410	-.457	.527	1			
Neg3	-.504	-.382	-.374	-.382	-.549	.639	.616	1		
Neg4	-.382	-.263	-.241	-.237	-.416	.469	.453	.465	1	
Neg5	-.519	-.366	-.368	-.401	-.557	.615	.538	.654	.534	1

Note. The table reports weighted statistics of the positively worded (Pos1–Pos5) and negatively worded (Neg1–Neg5) items that had a five-category Likert response scale (1 = *trifft gar nicht zu*, 2 = *trifft eher nicht zu*, 3 = *teils/teils*, 4 = *trifft eher zu*, 5 = *trifft völlig zu*). The correlations are based on pairwise complete observations.

The overall sample size amounted to $n = 4,791$ students.

Appendix 1.2

Item Wording, Descriptive Statistics, and Item Intercorrelations of the Mathematics Self-Concept Scale in PIRLS/TIMSS in Australia in Original Sequence (Top) and Ordered by Item Type (Bottom)

Items	How much do you agree with these statements about learning mathematics?	<i>M</i>	<i>SD</i>	Skewness	Excess kurtosis	Missing values
Pos1	I usually do well in mathematics.	1.64	0.81	1.25	1.02	1%
Neg1	Mathematics is harder for me than for many of my classmates.	2.96	1.08	-0.55	-1.07	2%
Neg2	I am just not good at mathematics.	3.16	1.05	-0.93	-0.50	3%
Pos2	I learn things quickly in mathematics.	1.84	0.93	0.89	-0.18	2%
Pos3	I am good at working out difficult mathematics problems.	2.05	1.00	0.61	-0.72	2%
Pos4	My teacher tells me I am good at mathematics.	1.93	0.95	0.77	-0.38	2%
Neg3	Mathematics is harder for me than any other subject.	3.04	1.13	-0.74	-0.95	2%

Item intercorrelations							
	Pos1	Pos2	Pos3	Pos4	Neg1	Neg2	Neg3
Pos1	1						
Pos2	0.605	1					
Pos3	0.597	0.624	1				
Pos4	0.516	0.508	0.494	1			
Neg1	-0.431	-0.401	-0.424	-0.303	1		
Neg2	-0.514	-0.470	-0.476	-0.356	0.622	1	
Neg3	-0.445	-0.418	-0.443	-0.308	0.594	0.585	1

Note. The table reports weighted statistics of the positively worded (Pos1–Pos4) and negatively worded (Neg1–Neg3) items that had a four-category Likert response scale (1 = *agree a lot*, 2 = *agree a little*, 3 = *disagree a little*, 4 = *disagree a lot*). The correlations are based on pairwise complete observations. The overall sample size amounted to $n = 5,943$ students.

Appendix 1.3

Item Wording, Descriptive Statistics, and Item Intercorrelations of the Reading Self-Concept Scale in PIRLS/TIMSS in Australia in Original Sequence (Top) and Ordered by Item Type (Bottom)

Items	How much do you agree with these statements about reading?	<i>M</i>	<i>SD</i>	Skew-ness	Excess kurtosis	Missing values
Pos1	I usually do well in reading.	1.50	0.72	1.53	2.26	2%
Pos2	Reading is easy for me.	1.53	0.76	1.48	1.84	2%
Neg1	Reading is harder for me than for many of my classmates.	3.16	1.00	-0.87	-0.51	3%
Pos3	If a book is interesting, I don't care how hard it is to read.	1.60	0.88	1.42	1.10	3%
Neg2	I have trouble reading stories with difficult words.	2.58	1.08	0.01	-1.29	3%
Pos4	My teacher tells me I am a good reader.	1.87	0.92	0.89	-0.05	3%
Neg3	Reading is harder for me than any other subject.	3.41	0.94	-1.42	0.76	3%

Item intercorrelations							
	Pos1	Pos2	Pos3	Pos4	Neg1	Neg2	Neg3
Pos1	1						
Pos2	0.589	1					
Pos3	0.332	0.311	1				
Pos4	0.310	0.223	0.192	1			
Neg1	-0.340	-0.448	-0.144	-0.083	1		
Neg2	-0.247	-0.340	-0.112	-0.049	0.431	1	
Neg3	-0.350	-0.404	-0.183	-0.094	0.497	0.374	1

Note. The table reports weighted statistics of the positively worded (Pos1–Pos4) and negatively worded (Neg1–Neg3) items that had a four-category Likert response scale (1 = *agree a lot*, 2 = *agree a little*, 3 = *disagree a little*, 4 = *disagree a lot*). The correlations are based on pairwise complete observations. The overall sample size amounted to $n = 5,943$ students.

Appendix 1.4

Item Wording, Descriptive Statistics, and Item Intercorrelations of the Reading Self-Concept Scale in PISA in the USA in Original Sequence (Top) and Ordered by Item Type (Bottom)

Items	How much do you agree or disagree with these statements about reading?	<i>M</i>	<i>SD</i>	Skew-ness	Excess kurtosis	Missing values
Pos1	I am a good reader	3.10	0.74	-0.67	0.48	2%
Pos2	I am able to understand difficult texts	2.90	0.76	-0.42	0.00	3%
Pos3	I read fluently	3.07	0.74	-0.53	0.11	3%
Neg1	I have always had difficulty with reading	1.92	0.84	0.66	-0.16	3%
Neg2	I have to read a text several times before completely understanding it	2.44	0.85	-0.03	-0.64	3%
Neg3	I find it difficult to answer questions about a text	2.11	0.80	0.45	-0.12	3%

Item intercorrelations						
	Pos1	Pos2	Pos3	Neg1	Neg2	Neg3
Pos1	1					
Pos2	.657	1				
Pos3	.690	.603	1			
Neg1	-.504	-.406	-.479	1		
Neg2	-.312	-.368	-.290	.495	1	
Neg3	-.374	-.389	-.341	.535	.530	1

Note. The table reports weighted statistics of the positively worded (Pos1–Pos3) and negatively worded (Neg1–Neg3) items that had a four-category Likert response scale (1 = *strongly disagree*, 2 = *disagree*, 3 = *agree*, 4 = *strongly agree*). The correlations are based on pairwise complete observations. The overall sample size amounted to $n = 4,799$ students.

Appendix 2: Mplus Syntaxes of the Analysis Models

Appendix 2.1

Mplus Syntax of the One-Dimensional CFA for the Self-Esteem Scale in NEPS Grade 5

```
TITLE:
NEPS Grade 5
Self-Esteem Scale
1-dimensional CFA;

DATA:
file = "neps5.dat";

VARIABLE:
names = ID school weight read speed reason
pos1 pos2 pos3 pos4 pos5 neg1 neg2 neg3 neg4 neg5;
usevar = pos1 pos2 pos3 pos4 pos5 neg1 neg2 neg3 neg4 neg5;
missing = .;
cluster = school;
weight = weight;

ANALYSIS:
type = complex;
estimator = MLR;

MODEL:
f by pos1 * pos2-pos5 (p1-p5)
neg1-neg5;
f@1;
[f@0];

MODEL CONSTRAINT:
DO (1, 5) p# > 0;
```


Appendix 2.2

Mplus Syntax of Two-Dimensional CFA for the Self-Esteem Scale in NEPS Grade 5

```
TITLE:
NEPS Grade 5
Self-Esteem Scale
2-dimensional CFA;

DATA:
file = "neps5.dat";

VARIABLE:
names = ID school weight read speed reason
pos1 pos2 pos3 pos4 pos5 neg1 neg2 neg3 neg4 neg5;
usevar = pos1 pos2 pos3 pos4 pos5 neg1 neg2 neg3 neg4 neg5;
missing = .;
cluster = school;
weight = weight;

ANALYSIS:
type = complex;
estimator = MLR;

MODEL:
f1 by pos1* pos2-pos5 (p1-p5)
neg1-neg5;
f1@1;
[f1@0];
f2 by neg1* neg2-neg5;
f2@1;
f1 with f2@0;

MODEL CONSTRAINT:
DO (1, 5) p# > 0;
```

Appendix 2.3

Mplus Syntax of Factor Mixture Analysis for the Self-Esteem Scale in NEPS Grade 5

```
TITLE:
NEPS Grade 5
Self-Esteem Scale
Factor Mixture Analysis;

DATA:
file = "neps5.dat";

VARIABLE:
names = ID school weight read speed reason
pos1 pos2 pos3 pos4 pos5 neg1 neg2 neg3 neg4 neg5;
usevar = pos1 pos2 pos3 pos4 pos5 neg1 neg2 neg3 neg4 neg5;
missing = .;
cluster = school;
weight = weight;
classes = c(2);

ANALYSIS:
type = mixture complex;
estimator = MLR;
starts = 5000 500;

MODEL:
%overall%
f by pos1* pos2-pos5 (p1-p5)
neg1-neg5;
f@1;
[f@0];
%c#1%
f;
[f];
f by neg1-neg5 (c1_n1-c1_n5);
[neg1-neg5];
%c#2%
f by neg1-neg5 (c2_n1-c2_n5);
[neg1-neg5];

MODEL CONSTRAINT:
DO (1, 5) p# > 0;
DO (1, 5) c1_n# > 0;
DO (1, 5) c1_n# = - c2_n#;

OUTPUT:
TECH7;
TECH12;

SAVEDATA:
file = neps5_fma.dat;
format = free;
save = CPROBABILITIES;
```

Appendix 2.4

Mplus Syntax of Factor Mixture Analysis for the Self-Esteem Scale in NEPS Grade 5 with Reading as Auxiliary

Variable

```

TITLE:
NEPS Grade 5
Self-Esteem Scale
Factor Mixture Analysis
with Auxiliary Reading;

DATA:
file = "neps5.dat";

VARIABLE:
names = ID school weight read speed reason
pos1 pos2 pos3 pos4 pos5 neg1 neg2 neg3 neg4 neg5;
usevar = pos1 pos2 pos3 pos4 pos5 neg1 neg2 neg3 neg4 neg5 read read_d;
auxiliary = read (R3STEP);
missing = .;
cluster = school;
weight = weight;
classes = c(2);

DEFINE:
read_d = read;
standardize read_d;

ANALYSIS:
type = mixture complex;
estimator = MLR;
starts = 5000 500;

MODEL:
%overall%
f by pos1* pos2-pos5 (p1-p5)
neg1-neg5;
f@1;
[f@0];
c on read_d;
%c#1%
f;
[f];
f by neg1-neg5 (c1_n1-c1_n5);
[neg1-neg5];
%c#2%
f by neg1-neg5 (c2_n1-c2_n5);
[neg1-neg5];

MODEL CONSTRAINT:
DO (1, 5) p# > 0;
DO (1, 5) c1_n# > 0;
DO (1, 5) c1_n# = - c2_n#;

```

Appendix 3: Model Results of the Generalization Analyses

Appendix 3.1

Model Estimation Results of the One-Dimensional (1-dim) and Two-Dimensional (2-dim) Confirmatory Factor Analyses (CFA) and Constrained Factor Mixture Analysis (FMA) for the Self-Esteem Scale in NEPS Grade 9 in Germany

	1-dim CFA		2-dim CFA				FMA			
	Global factor		Global factor		Method factor		Consistent class 93%		Inconsistent class 7%	
	Par.	SE	Par.	SE	Par.	SE	Par.	SE	Par.	SE
<i>Factors</i>										
Mean α	0		0		0		0		-0.93	0.16
Variance ϕ	1		1		1		1		1.13	0.15
<i>Loadings λ</i>										
Pos1	0.69	0.03	0.73	0.03			0.69	0.03	0.69	0.03
Pos2	0.50	0.03	0.58	0.03			0.53	0.03	0.53	0.03
Pos3	0.49	0.03	0.56	0.03			0.52	0.03	0.52	0.03
Pos4	0.62	0.03	0.68	0.03			0.63	0.03	0.63	0.03
Pos5	0.78	0.03	0.84	0.02			0.80	0.03	0.80	0.03
Neg1	-0.80	0.03	-0.63	0.03	0.58	0.03	-0.77	0.03	0.77	0.03
Neg2	-0.76	0.03	-0.63	0.03	0.46	0.04	-0.77	0.03	0.77	0.03
Neg3	-0.90	0.03	-0.73	0.04	0.61	0.04	-0.90	0.03	0.90	0.03
Neg4	-0.67	0.03	-0.52	0.03	0.47	0.04	-0.64	0.03	0.64	0.03
Neg5	-0.86	0.03	-0.71	0.03	0.53	0.04	-0.78	0.04	0.78	0.04
<i>Intercepts ν</i>										
Pos1	3.77	0.03	3.77	0.03			3.82	0.03	3.82	0.03
Pos2	3.82	0.03	3.82	0.03			3.86	0.03	3.86	0.03
Pos3	3.80	0.03	3.80	0.03			3.83	0.03	3.83	0.03
Pos4	3.68	0.03	3.68	0.03			3.72	0.03	3.72	0.03
Pos5	3.67	0.03	3.67	0.03			3.72	0.03	3.72	0.03
Neg1	2.48	0.03	2.48	0.03			2.42	0.03	4.02	0.18
Neg2	2.26	0.03	2.26	0.03			2.21	0.03	3.66	0.15
Neg3	2.12	0.03	2.12	0.03			2.06	0.04	3.86	0.15
Neg4	2.61	0.03	2.61	0.03			2.55	0.04	4.03	0.19
Neg5	1.99	0.03	1.99	0.03			1.88	0.04	4.19	0.26
<i>Res. var. Ψ</i>										
Pos1	0.48	0.04	0.42	0.04			0.45	0.03	0.45	0.03
Pos2	0.48	0.02	0.39	0.02			0.42	0.02	0.42	0.02
Pos3	0.50	0.03	0.42	0.03			0.45	0.02	0.45	0.02
Pos4	0.76	0.05	0.67	0.05			0.71	0.05	0.71	0.05
Pos5	0.47	0.04	0.38	0.03			0.41	0.03	0.41	0.03
Neg1	0.60	0.03	0.51	0.03			0.59	0.03	0.59	0.03
Neg2	0.62	0.05	0.59	0.04			0.57	0.04	0.57	0.04
Neg3	0.51	0.04	0.42	0.04			0.45	0.04	0.45	0.04
Neg4	0.86	0.04	0.81	0.04			0.84	0.03	0.84	0.03
Neg5	0.49	0.03	0.44	0.04			0.45	0.03	0.45	0.03

Note. The positively worded (Pos1–Pos5) and negatively worded (Neg1–Neg5) indicators had a five-category Likert response scale (1 = *does not apply at all*, 2 = *does rather not apply*, 3 = *partly*, 4 = *does rather apply*, 5 = *applies completely*). The table reports unstandardized parameter estimates and standard errors of model estimations in which sampling weights were applied. Fixed latent means and variances are depicted without standard errors. For the two-dimensional CFA model, the parameters for the second and orthogonal method factor are displayed in a separate column. For the constrained FMA model, the parameters are displayed for the two classes of consistent versus inconsistent respondents separately. The reported class sizes are the final class proportions based on the estimated model in percent. Cases with missing values on all variables were excluded from the analyses. The effective sample size amounted to $n = 4,754$ students.

Appendix 3.2

Model Estimation Results of the One-Dimensional (1-dim) and Two-Dimensional (2-dim) Confirmatory Factor Analyses (CFA) and Constrained Factor Mixture Analysis (FMA) for the Mathematics Self-Concept Scale in PIRLS/TIMSS in Australia

	1-dim CFA		2-dim CFA				FMA			
	Global factor		Global factor		Method factor		Consistent class 88%		Inconsistent class 12%	
	Par.	SE	Par.	SE	Par.	SE	Par.	SE	Par.	SE
<i>Factors</i>										
Mean α	0		0		0		0		0.32	0.08
Variance ϕ	1		1		1		1		0.90	0.07
<i>Loadings λ</i>										
Pos1	0.62	0.01	0.64	0.01			0.65	0.02	0.65	0.02
Pos2	0.69	0.02	0.73	0.02			0.72	0.02	0.72	0.02
Pos3	0.75	0.02	0.78	0.02			0.76	0.02	0.76	0.02
Pos4	0.57	0.02	0.61	0.02			0.59	0.02	0.59	0.02
Neg1	-0.70	0.02	-0.57	0.02	0.66	0.03	-0.72	0.02	0.72	0.02
Neg2	-0.75	0.02	-0.65	0.02	0.52	0.03	-0.81	0.02	0.81	0.02
Neg3	-0.74	0.02	-0.62	0.02	0.57	0.03	-0.74	0.02	0.74	0.02
<i>Intercepts ν</i>										
Pos1	1.64	0.02	1.64	0.02			1.61	0.02	1.61	0.02
Pos2	1.84	0.02	1.84	0.02			1.81	0.02	1.81	0.02
Pos3	2.04	0.02	2.04	0.02			2.01	0.02	2.01	0.02
Pos4	1.93	0.02	1.93	0.02			1.90	0.02	1.90	0.02
Neg1	2.96	0.02	2.96	0.02			3.06	0.02	1.99	0.10
Neg2	3.17	0.02	3.17	0.02			3.32	0.03	1.77	0.08
Neg3	3.04	0.02	3.04	0.02			3.13	0.03	2.17	0.13
<i>Res. var. Ψ</i>										
Pos1	0.28	0.02	0.25	0.02			0.25	0.02	0.25	0.02
Pos2	0.39	0.02	0.34	0.02			0.35	0.02	0.35	0.02
Pos3	0.44	0.02	0.39	0.02			0.41	0.02	0.41	0.02
Pos4	0.58	0.02	0.54	0.02			0.56	0.02	0.56	0.02
Neg1	0.68	0.03	0.42	0.03			0.58	0.03	0.58	0.03
Neg2	0.54	0.03	0.42	0.02			0.27	0.02	0.27	0.02
Neg3	0.73	0.04	0.57	0.03			0.68	0.03	0.68	0.03

Note. The positively worded (Pos1–Pos4) and negatively worded (Neg1–Neg3) indicators had a four-category

Likert response scale (1 = *agree a lot*, 2 = *agree a little*, 3 = *disagree a little*, 4 = *disagree a lot*). The table reports unstandardized parameter estimates and standard errors of model estimations in which sampling weights were applied. Fixed latent means and variances are depicted without standard errors. For the two-dimensional CFA model, the parameters for the second and orthogonal method factor are displayed in a separate column. For the constrained FMA model, the parameters are displayed for the two classes of consistent versus inconsistent respondents separately. The reported class sizes are the final class proportions based on the estimated model in

percent. Cases with missing values on all variables were excluded from the analyses. The effective sample size amounted to $n = 5,885$ students.

Appendix 3.3

Model Estimation Results of the One-Dimensional (1-dim) and Two-Dimensional (2-dim) Confirmatory Factor Analyses (CFA) and Constrained Factor Mixture Analysis (FMA) for the Reading Self-Concept Scale in PIRLS/TIMSS in Australia

	1-dim CFA		2-dim CFA				FMA			
	Global factor		Global factor		Method factor		Consistent class 80%		Inconsistent class 20%	
	Par.	SE	Par.	SE	Par.	SE	Par.	SE	Par.	SE
<i>Factors</i>										
Mean α	0		0		0		0		2.04	0.46
Variance ϕ	1		1		1		1		3.99	1.35
<i>Loadings λ</i>										
Pos1	0.50	0.02	0.53	0.02			0.37	0.05	0.37	0.05
Pos2	0.59	0.02	0.62	0.02			0.40	0.05	0.40	0.05
Pos3	0.33	0.03	0.36	0.03			0.25	0.03	0.25	0.03
Pos4	0.26	0.02	0.29	0.02			0.20	0.04	0.20	0.04
Neg1	-0.61	0.03	-0.50	0.03	0.57	0.03	-0.17	0.07	0.17	0.07
Neg2	-0.51	0.02	-0.40	0.02	0.46	0.03	-0.21	0.06	0.21	0.06
Neg3	-0.55	0.02	-0.45	0.02	0.43	0.03	-0.14	0.02	0.14	0.02
<i>Intercepts ν</i>										
Pos1	1.50	0.01	1.50	0.01			1.35	0.02	1.35	0.02
Pos2	1.53	0.02	1.53	0.02			1.37	0.03	1.37	0.03
Pos3	1.60	0.02	1.60	0.02			1.50	0.02	1.50	0.02
Pos4	1.87	0.02	1.87	0.02			1.79	0.03	1.79	0.03
Neg1	3.16	0.02	3.16	0.02			3.39	0.03	1.88	0.10
Neg2	2.58	0.02	2.59	0.02			2.76	0.03	1.46	0.08
Neg3	3.41	0.03	3.41	0.03			3.81	0.01	1.49	0.04
<i>Res. var. Ψ</i>										
Pos1	0.27	0.01	0.23	0.01			0.21	0.02	0.21	0.02
Pos2	0.23	0.02	0.20	0.01			0.23	0.02	0.23	0.02
Pos3	0.66	0.02	0.65	0.02			0.63	0.02	0.63	0.02
Pos4	0.78	0.03	0.76	0.03			0.75	0.03	0.75	0.03
Neg1	0.64	0.03	0.43	0.03			0.75	0.03	0.75	0.03
Neg2	0.89	0.03	0.78	0.03			0.97	0.03	0.97	0.03
Neg3	0.58	0.03	0.50	0.03			0.20	0.02	0.20	0.02

Note. The positively worded (Pos1–Pos4) and negatively worded (Neg1–Neg3) indicators had a four-category

Likert response scale (1 = *agree a lot*, 2 = *agree a little*, 3 = *disagree a little*, 4 = *disagree a lot*). The table reports unstandardized parameter estimates and standard errors of model estimations in which sampling weights were applied. Fixed latent means and variances are depicted without standard errors. For the two-dimensional CFA model, the parameters for the second and orthogonal method factor are displayed in a separate column. For the constrained FMA model, the parameters are displayed for the two classes of consistent versus inconsistent respondents separately. The reported class sizes are the final class proportions based on the estimated model in

percent. Cases with missing values on all variables were excluded from the analyses. The effective sample size amounted to $n = 5,863$ students.

Appendix 3.4

Model Estimation Results of the One-Dimensional (1-dim) and Two-Dimensional (2-dim) Confirmatory Factor Analyses (CFA) and Constrained Factor Mixture Analysis (FMA) for the Reading Self-Concept Scale in PISA in the USA

	1-dim CFA		2-dim CFA				FMA			
	Global factor		Global factor		Method factor		Consistent class 92%		Inconsistent class 8%	
	Par.	SE	Par.	SE	Par.	SE	Par.	SE	Par.	SE
<i>Factors</i>										
Mean α	0		0		0		0		-0.72	0.10
Variance ϕ	1		1		1		1		1.72	0.16
<i>Loadings λ</i>										
Pos1	0.61	0.01	0.64	0.01			0.59	0.01	0.59	0.01
Pos2	0.57	0.01	0.57	0.01			0.56	0.01	0.56	0.01
Pos3	0.58	0.01	0.59	0.01			0.55	0.01	0.55	0.01
Neg1	-0.54	0.02	-0.48	0.02	0.38	0.02	-0.59	0.01	0.59	0.01
Neg2	-0.41	0.02	-0.33	0.02	0.50	0.02	-0.49	0.01	0.49	0.01
Neg3	-0.43	0.02	-0.36	0.02	0.48	0.02	-0.49	0.01	0.49	0.01
<i>Intercepts ν</i>										
Pos1	3.11	0.01	3.11	0.01			3.14	0.01	3.14	0.01
Pos2	2.91	0.02	2.91	0.02			2.94	0.02	2.94	0.02
Pos3	3.08	0.02	3.08	0.02			3.11	0.02	3.11	0.02
Neg1	1.91	0.02	1.91	0.02			1.85	0.02	3.03	0.08
Neg2	2.44	0.02	2.44	0.02			2.41	0.02	3.10	0.05
Neg3	2.10	0.01	2.10	0.01			2.07	0.02	2.89	0.06
<i>Res. var. Ψ</i>										
Pos1	0.17	0.01	0.14	0.01			0.17	0.01	0.17	0.01
Pos2	0.24	0.01	0.24	0.01			0.24	0.01	0.24	0.01
Pos3	0.21	0.01	0.20	0.01			0.21	0.01	0.21	0.01
Neg1	0.41	0.02	0.32	0.01			0.28	0.01	0.28	0.01
Neg2	0.55	0.02	0.36	0.02			0.46	0.01	0.46	0.01
Neg3	0.45	0.02	0.28	0.02			0.36	0.01	0.36	0.01

Note. The positively worded (Pos1–Pos3) and negatively worded (Neg1–Neg3) indicators had a four-category

Likert response scale (1 = *strongly disagree*, 2 = *disagree*, 3 = *agree*, 4 = *strongly agree*). The table reports unstandardized parameter estimates and standard errors of model estimations in which sampling weights were applied. Fixed latent means and variances are depicted without standard errors. For the two-dimensional CFA model, the parameters for the second and orthogonal method factor are displayed in a separate column. For the constrained FMA model, the parameters are displayed for the two classes of consistent versus inconsistent respondents separately. The reported class sizes are the final class proportions based on the estimated model in percent. Cases with missing values on all variables were excluded from the analyses. The effective sample size amounted to $n = 4,703$ students.

Appendix 4: Item Statistics within Latent Classes

Appendix 4.1

Average Estimated Means, Standard Deviations, and Item Intercorrelations across Positively and Negatively Worded Items in the Latent Classes of the Constrained Factor Mixture Analysis (FMA) for the Self-Esteem Scale in NEPS Grade 9 in Germany

	Consistent class 93%		Inconsistent class 7%	
	Average Pos1–Pos5	Average Neg1–Neg5	Average Pos1–Pos5	Average Neg1–Neg5
<i>M</i>	3.79	2.22	3.20	3.23
<i>SD</i>	0.95	1.09	0.97	1.12
Item intercorrelation				
Average Pos1–Pos5	.446		.476	
Average Neg1–Neg5	-.476	.504	.506	.534

Note. The positively worded (Pos1–Pos5) and negatively worded (Neg1–Neg5) indicators had a five-category

Likert response scale (1 = *does not apply at all*, 2 = *does rather not apply*, 3 = *partly*, 4 = *does rather apply*, 5 = *applies completely*). The table reports average estimates of model estimations in which sampling weights were applied. Cases with missing values on all variables were excluded from the analyses. The effective sample size amounted to $n = 4,754$ students.

Appendix 4.2

Average Estimated Means, Standard Deviations, and Item Intercorrelations across Positively and Negatively Worded Items in the Latent Classes of the Constrained Factor Mixture Analysis (FMA) for the Mathematics Self-Concept Scale in PIRLS/TIMSS in Australia

	Consistent class 88%		Inconsistent class 12%	
	Average Pos1–Pos4	Average Neg1–Neg3	Average Pos1–Pos4	Average Neg1–Neg3
<i>M</i>	1.83	3.17	2.05	2.22
<i>SD</i>	0.92	1.04	0.90	1.01
Item intercorrelation				
Average Pos1–Pos4	.541		.516	
Average Neg1–Neg3	-.540	.534	.515	.509

Note. The positively worded (Pos1–Pos4) and negatively worded (Neg1–Neg3) indicators had a four-category

Likert response scale (1 = *agree a lot*, 2 = *agree a little*, 3 = *disagree a little*, 4 = *disagree a lot*). The table reports average estimates of model estimations in which sampling weights were applied. Cases with missing values on all variables were excluded from the analyses. The effective sample size amounted to $n = 5,885$ students.

Appendix 4.3

Average Estimated Means, Standard Deviations, and Item Intercorrelations across Positively and Negatively Worded Items in the Latent Classes of the Constrained Factor Mixture Analysis (FMA) for the Reading Self-Concept Scale in PIRLS/TIMSS in Australia

	Consistent class 80%		Inconsistent class 20%	
	Average Pos1–Pos4	Average Neg1–Neg3	Average Pos1–Pos4	Average Neg1–Neg3
<i>M</i>	1.50	3.32	2.13	1.96
<i>SD</i>	0.73	0.78	0.92	0.84
Item intercorrelation				
Average Pos1–Pos4	.189		.430	
Average Neg1–Neg3	-.103	.052	.282	.177

Note. The positively worded (Pos1–Pos4) and negatively worded (Neg1–Neg3) indicators had a four-category

Likert response scale (1 = *agree a lot*, 2 = *agree a little*, 3 = *disagree a little*, 4 = *disagree a lot*). The table reports average estimates of model estimations in which sampling weights were applied. Cases with missing values on all variables were excluded from the analyses. The effective sample size amounted to $n = 5,863$ students.

Appendix 4.4

Average Estimated Means, Standard Deviations, and Item Intercorrelations across Positively and Negatively Worded Items in the Latent Classes of the Constrained Factor Mixture Analysis (FMA) for the Reading Self-Concept Scale in PISA in the USA

	Consistent class 92%		Inconsistent class 8%	
	Average Pos1–Pos3	Average Neg1–Neg3	Average Pos1–Pos3	Average Neg1–Neg3
<i>M</i>	3.06	2.11	2.65	2.63
<i>SD</i>	0.72	0.80	0.87	0.92
Item intercorrelation				
Average Pos1–Pos3	.612		.730	
Average Neg1–Neg3	-.512	.425	.639	.558

Note. The positively worded (Pos1–Pos3) and negatively worded (Neg1–Neg3) indicators had a four-category

Likert response scale (1 = *strongly disagree*, 2 = *disagree*, 3 = *agree*, 4 = *strongly agree*). The table reports

average estimates of model estimations in which sampling weights were applied. Cases with missing values on

all variables were excluded from the analyses. The effective sample size amounted to $n = 4,703$ students.

Appendix 5: Less Constrained FMA Models

We ran two less constrained FMA models to explore the necessity of the model constraints and to evaluate the model fit (cf. Appendix 5.1). At first, we fitted a fully exploratory configural invariant model (i.e., with class-noninvariant intercepts and loadings for all items) to serve as a best-fit reference for any two-class FMA model (hereafter non-invariant FMA). Second, we fitted a FMA that was less constrained than the main model but more constrained than the non-invariant model. We loosened the constrained FMA's restrictions of the factor loadings of the negatively worded items (hereafter neg. invariant FMA). The aim of this additional analysis was to explore the consequences of relaxing the assumption of polar opposite class invariance of the negatively worded items.

Both types of less constrained models resulted in one larger class with opposite loadings depending on the item wording (i.e., a consistently responding class) for all five datasets. The factor loading patterns of the smaller classes differed, however. In three cases, the smaller class showed the same factor loading signs for positively and negatively worded items and therefore resembled the inconsistently responding class in the constrained FMA (neg. invariant FMA in Appendix 5.3, neg. invariant and non-invariant FMA in Appendix 5.6). In four cases, the smaller class showed opposite factor loadings for the positively and negatively worded items (i.e., consistent responses), but the factor loadings of the negatively worded items were closer to zero (neg. invariant and non-invariant FMA in Appendixes 5.4 and 5.5). In the remaining three cases, the smaller classes showed close to zero factor variances in combination with out-of-boundary loadings and intercepts (neg. invariant FMA in Appendix 5.2), or only significant factor loadings of (most) negatively worded items and insignificant factor loadings of (most) positively worded items (non-invariant FMAs in Appendixes 5.2 and 5.3). In summary, the less constrained FMAs picked up on class distinctions as a function of differential responding depending on item wording, but not always in the way that was intended in this study (i.e., consistent and inconsistent respondents to mixed-worded scales).

When evaluating the model fit, clear patterns can be observed (see Appendix 5.1). In the cases where the less constrained models resulted in similar classes as the constrained FMAs (i.e., consistently and inconsistently responding classes) and which are therefore substantive competitors, the constrained FMAs showed an at least equally good model fit as the less constrained models. This is an important finding in support of our main models. If the less constrained models did not result in similar classifications as the constrained FMAs, the less constrained models showed a superior model fit, despite the penalties for additional parameters. However, these models did not identify the two theoretical latent classes of interest for the present study, but rather picked up on other unmodeled data features. This further confirms the necessity of the strict model

constraints in the main FMA models to identify the latent classes of interest. Furthermore, in comparison with the fit of the non-invariant FMAs, the fit of the neg. invariant FMAs was about the same or only slightly inferior in all five datasets. This provides further support for the tenability of the strong constraints on the positively worded items in our constrained model.

Appendix 5.1

Model Fit Results of the Constrained Factor Mixture Analysis (Constrained FMA), the Factor Mixture Analysis with Invariant Negatively Worded Items (Neg. Invariant FMA), and the Non-Invariant Factor Mixture Analysis (Non-Invariant FMA)

	Constrained FMA	Neg. invariant FMA	Non-invariant FMA
<i>Self-esteem NEPS grade 5</i>			
<i>n</i>	4,938	4,938	4,938
No. par.	38	43	51
AIC	135,119	134,324	134,299
BIC	135,367	134,604	134,631
Entropy	.831	.913	.912
<i>Self-esteem NEPS grade 9</i>			
<i>n</i>	4,754	4,754	4,754
No. par.	38	43	51
AIC	115,873	115,808	115,639
BIC	116,118	116,086	115,969
Entropy	.882	.887	.871
<i>Math. self-concept PIRLS/TIMSS</i>			
<i>n</i>	5,885	5,885	5,885
No. par.	27	30	36
AIC	96,490	94,913	94,811
BIC	96,670	95,113	95,051
Entropy	.810	.963	.960
<i>Read. self-concept PIRLS/TIMSS</i>			
<i>n</i>	5,863	5,863	5,863
No. par.	27	30	36
AIC	94,364	93,262	93,226
BIC	94,545	93,463	93,467
Entropy	.954	.988	.988
<i>Read. self-concept PISA</i>			
<i>n</i>	4,703	4,703	4,703
No. par.	24	27	31
AIC	54,272	54,273	54,233
BIC	54,427	54,447	54,433
Entropy	.814	.813	.818

Note. The model fit results of the constrained FMAs are identical to those reported in Table 5.

Appendix 5.2

Model Estimation Results of the Factor Mixture Analysis with Invariant Negatively Worded Items (Neg. Invariant FMA), and the Non-Invariant Factor Mixture Analysis (Non-Invariant FMA) for the Self-Esteem Scale in NEPS Grade 5 in Germany

	Neg. invariant FMA				Non-invariant FMA			
	C#1 85%		C#2 15%		C#1 85%		C#2 15%	
	Par.	SE	Par.	SE	Par.	SE	Par.	SE
<i>Factors</i>								
Mean α	0		-0.55	0.09	0		0	
Variance ϕ	1		0.00	0.01	1		1	
<i>Loadings λ</i>								
Pos1	-0.58	0.02	-0.58	0.02	-0.58	0.02	-0.05	0.06
Pos2	-0.52	0.02	-0.52	0.02	-0.52	0.03	-0.02	0.05
Pos3	-0.54	0.02	-0.54	0.02	-0.54	0.02	0.01	0.06
Pos4	-0.59	0.03	-0.59	0.03	-0.60	0.03	0.03	0.10
Pos5	-0.68	0.02	-0.68	0.02	-0.68	0.02	-0.06	0.10
Neg1	0.67	0.02	22.48	56.14	0.67	0.03	0.90	0.06
Neg2	0.71	0.03	-2.56	5.33	0.70	0.03	-0.10	0.07
Neg3	0.75	0.03	32.28	80.03	0.75	0.03	1.30	0.04
Neg4	0.42	0.03	16.80	42.44	0.42	0.03	0.67	0.07
Neg5	0.70	0.03	30.10	74.38	0.70	0.03	1.21	0.04
<i>Intercepts ν</i>								
Pos1	4.24	0.02	4.24	0.02	4.24	0.02	4.55	0.06
Pos2	4.03	0.02	4.03	0.02	4.02	0.02	4.43	0.05
Pos3	3.97	0.02	3.97	0.02	3.96	0.02	4.32	0.06
Pos4	3.85	0.03	3.85	0.03	3.87	0.03	4.03	0.06
Pos5	3.98	0.02	3.98	0.02	3.99	0.02	4.27	0.08
Neg1	2.35	0.03	15.44	32.40	2.35	0.03	3.00	0.07
Neg2	1.81	0.03	2.94	3.06	1.81	0.03	4.36	0.07
Neg3	1.73	0.03	20.34	46.18	1.73	0.03	2.47	0.08
Neg4	2.60	0.03	12.83	24.52	2.60	0.03	3.52	0.10
Neg5	1.69	0.02	19.08	42.94	1.69	0.02	2.41	0.07
<i>Res. var. Ψ</i>								
Pos1	0.45	0.02	0.45	0.02	0.45	0.02	0.45	0.02
Pos2	0.57	0.02	0.57	0.02	0.57	0.02	0.57	0.02
Pos3	0.67	0.02	0.67	0.02	0.66	0.02	0.66	0.02
Pos4	1.24	0.05	1.24	0.05	1.23	0.04	1.23	0.04
Pos5	0.71	0.03	0.71	0.03	0.71	0.03	0.71	0.03
Neg1	1.17	0.03	1.17	0.03	1.17	0.03	1.17	0.03
Neg2	0.47	0.03	0.47	0.03	0.47	0.03	0.47	0.03
Neg3	0.58	0.03	0.58	0.03	0.58	0.03	0.58	0.03
Neg4	1.54	0.04	1.54	0.04	1.55	0.04	1.55	0.04
Neg5	0.62	0.04	0.62	0.04	0.62	0.04	0.62	0.04

Note. The positively worded (Pos1–Pos5) and negatively worded (Neg1–Neg5) indicators had a five-category

Likert response scale (1 = *does not apply at all*, 2 = *does rather not apply*, 3 = *partly*, 4 = *does rather apply*, 5 = *applies completely*). The table reports unstandardized parameter estimates and standard errors of model

estimations in which sampling weights were applied. Fixed latent means and variances are depicted without standard errors. The reported class sizes are the final class proportions based on the estimated model in percent. Cases with missing values on all variables were excluded from the analyses. The effective sample size amounted to $n = 4,938$ students.

Appendix 5.3

Model Estimation Results of the Factor Mixture Analysis with Invariant Negatively Worded Items (Neg. Invariant FMA), and the Non-Invariant Factor Mixture Analysis (Non-Invariant FMA) for the Self-Esteem Scale in NEPS Grade 9 in Germany

	Neg. invariant FMA				Non-invariant FMA			
	C#1 93%		C#2 7%		C#1 91%		C#2 9%	
	Par.	SE	Par.	SE	Par.	SE	Par.	SE
<i>Factors</i>								
Mean α	0		0.88	0.16	0		0	
Variance ϕ	1		1.50	0.19	1		1	
<i>Loadings λ</i>								
Pos1	-0.69	0.03	-0.69	0.03	0.74	0.03	0.05	0.10
Pos2	-0.53	0.03	-0.53	0.03	0.55	0.03	-0.24	0.15
Pos3	-0.52	0.03	-0.52	0.03	0.53	0.03	-0.26	0.16
Pos4	-0.63	0.03	-0.63	0.03	0.72	0.03	-1.05	0.13
Pos5	-0.80	0.03	-0.80	0.03	0.84	0.03	-0.14	0.14
Neg1	0.78	0.03	-0.49	0.10	-0.79	0.03	-0.97	0.12
Neg2	0.76	0.03	-0.73	0.08	-0.75	0.03	-0.96	0.16
Neg3	0.90	0.03	-0.79	0.08	-0.88	0.04	-1.13	0.12
Neg4	0.64	0.03	-0.55	0.12	-0.65	0.03	-1.01	0.10
Neg5	0.80	0.03	-0.45	0.18	-0.85	0.03	-1.05	0.10
<i>Intercepts ν</i>								
Pos1	3.81	0.03	3.81	0.03	3.73	0.03	4.19	0.13
Pos2	3.85	0.03	3.85	0.03	3.81	0.03	4.00	0.22
Pos3	3.83	0.03	3.83	0.03	3.79	0.03	3.94	0.23
Pos4	3.72	0.03	3.72	0.03	3.78	0.04	2.68	0.20
Pos5	3.71	0.03	3.71	0.03	3.65	0.04	3.84	0.16
Neg1	2.42	0.03	3.77	0.14	2.46	0.03	2.71	0.15
Neg2	2.21	0.03	3.59	0.21	2.20	0.04	2.85	0.22
Neg3	2.06	0.04	3.74	0.17	2.08	0.04	2.53	0.22
Neg4	2.55	0.04	3.93	0.19	2.61	0.04	2.66	0.13
Neg5	1.88	0.04	3.86	0.29	1.97	0.04	2.21	0.18
<i>Res. var. Ψ</i>								
Pos1	0.44	0.03	0.44	0.03	0.44	0.03	0.44	0.03
Pos2	0.42	0.02	0.42	0.02	0.44	0.02	0.44	0.02
Pos3	0.45	0.02	0.45	0.02	0.47	0.03	0.47	0.03
Pos4	0.71	0.05	0.71	0.05	0.48	0.03	0.48	0.03
Pos5	0.41	0.03	0.41	0.03	0.43	0.03	0.43	0.03
Neg1	0.59	0.03	0.59	0.03	0.59	0.03	0.59	0.03
Neg2	0.56	0.04	0.56	0.04	0.57	0.06	0.57	0.06
Neg3	0.44	0.04	0.44	0.04	0.49	0.05	0.49	0.05
Neg4	0.84	0.03	0.84	0.03	0.83	0.04	0.83	0.04
Neg5	0.46	0.04	0.46	0.04	0.46	0.04	0.46	0.04

Note. The positively worded (Pos1–Pos5) and negatively worded (Neg1–Neg5) indicators had a five-category

Likert response scale (1 = *does not apply at all*, 2 = *does rather not apply*, 3 = *partly*, 4 = *does rather apply*, 5 = *applies completely*). The table reports unstandardized parameter estimates and standard errors of model

estimations in which sampling weights were applied. Fixed latent means and variances are depicted without standard errors. The reported class sizes are the final class proportions based on the estimated model in percent. Cases with missing values on all variables were excluded from the analyses. The effective sample size amounted to $n = 4,754$ students.

Appendix 5.4

Model Estimation Results of the Factor Mixture Analysis with Invariant Negatively Worded Items (Neg. Invariant FMA), and the Non-Invariant Factor Mixture Analysis (Non-Invariant FMA) for the Mathematics Self-Concept Scale in PIRLS/TIMSS in Australia

	Neg. invariant FMA				Non-invariant FMA			
	C#1 75%		C#2 25%		C#1 75%		C#2 25%	
	Par.	SE	Par.	SE	Par.	SE	Par.	SE
<i>Factors</i>								
Mean α	0		1.56	0.08	0		0	
Variance ϕ	1		2.18	0.14	1		1	
<i>Loadings λ</i>								
Pos1	0.49	0.01	0.49	0.01	-0.44	0.01	-0.76	0.02
Pos2	0.54	0.01	0.54	0.01	-0.56	0.02	-0.82	0.03
Pos3	0.58	0.02	0.58	0.02	-0.61	0.02	-0.78	0.03
Pos4	0.45	0.01	0.45	0.01	-0.50	0.02	-0.67	0.03
Neg1	-0.43	0.02	-0.15	0.03	0.43	0.02	0.22	0.04
Neg2	-0.25	0.01	-0.09	0.01	0.25	0.01	0.13	0.02
Neg3	-0.47	0.02	-0.19	0.04	0.46	0.02	0.27	0.06
<i>Intercepts ν</i>								
Pos1	1.45	0.01	1.45	0.01	1.43	0.01	2.24	0.03
Pos2	1.62	0.02	1.62	0.02	1.63	0.02	2.44	0.04
Pos3	1.81	0.02	1.81	0.02	1.81	0.02	2.73	0.04
Pos4	1.75	0.02	1.75	0.02	1.77	0.02	2.38	0.04
Neg1	3.31	0.02	2.16	0.07	3.31	0.02	1.93	0.04
Neg2	3.71	0.01	1.70	0.04	3.71	0.01	1.56	0.02
Neg3	3.39	0.02	2.30	0.09	3.39	0.02	2.00	0.04
<i>Res. var. Ψ</i>								
Pos1	0.25	0.01	0.25	0.01	0.25	0.02	0.25	0.02
Pos2	0.35	0.01	0.35	0.01	0.34	0.02	0.34	0.02
Pos3	0.41	0.02	0.41	0.02	0.41	0.02	0.41	0.02
Pos4	0.55	0.02	0.55	0.02	0.54	0.02	0.54	0.02
Neg1	0.66	0.02	0.66	0.02	0.66	0.02	0.66	0.02
Neg2	0.17	0.00	0.17	0.00	0.18	0.00	0.18	0.00
Neg3	0.74	0.03	0.74	0.03	0.74	0.03	0.74	0.03

Note. The positively worded (Pos1–Pos4) and negatively worded (Neg1–Neg3) indicators had a four-category

Likert response scale (1 = *agree a lot*, 2 = *agree a little*, 3 = *disagree a little*, 4 = *disagree a lot*). The table

reports unstandardized parameter estimates and standard errors of model estimations in which sampling weights

were applied. Fixed latent means and variances are depicted without standard errors. The reported class sizes are

the final class proportions based on the estimated model in percent. Cases with missing values on all variables

were excluded from the analyses. The effective sample size amounted to $n = 5,885$ students.

Appendix 5.5

Model Estimation Results of the Factor Mixture Analysis with Invariant Negatively Worded Items (Neg. Invariant FMA), and the Non-Invariant Factor Mixture Analysis (Non-Invariant FMA) for the Reading Self-Concept Scale in PIRLS/TIMSS in Australia

	Neg. invariant FMA				Non-invariant FMA			
	C#1 82%		C#2 18%		C#1 82%		C#2 18%	
	Par.	SE	Par.	SE	Par.	SE	Par.	SE
<i>Factors</i>								
Mean α	0		-1.40	0.12	0		0	
Variance ϕ	1		2.81	0.35	1		1	
<i>Loadings λ</i>								
Pos1	-0.41	0.01	-0.41	0.01	-0.41	0.01	-0.70	0.04
Pos2	-0.49	0.02	-0.49	0.02	-0.47	0.02	-0.80	0.05
Pos3	-0.27	0.02	-0.27	0.02	-0.28	0.02	-0.45	0.08
Pos4	-0.22	0.03	-0.22	0.03	-0.28	0.02	-0.35	0.09
Neg1	0.45	0.02	0.15	0.04	0.45	0.02	0.25	0.07
Neg2	0.44	0.03	0.06	0.03	0.44	0.03	0.10	0.06
Neg3	0.16	0.01	0.04	0.02	0.16	0.01	0.06	0.03
<i>Intercepts ν</i>								
Pos1	1.40	0.01	1.40	0.01	1.40	0.01	1.96	0.04
Pos2	1.41	0.01	1.41	0.01	1.41	0.01	2.11	0.04
Pos3	1.53	0.02	1.53	0.02	1.53	0.02	1.90	0.05
Pos4	1.81	0.03	1.81	0.03	1.84	0.03	1.99	0.04
Neg1	3.36	0.02	2.44	0.07	3.36	0.02	2.23	0.04
Neg2	2.74	0.02	1.94	0.06	2.74	0.02	1.85	0.04
Neg3	3.80	0.01	1.64	0.04	3.80	0.01	1.59	0.03
<i>Res. var. Ψ</i>								
Pos1	0.25	0.01	0.25	0.01	0.24	0.01	0.24	0.01
Pos2	0.20	0.01	0.20	0.01	0.21	0.01	0.21	0.01
Pos3	0.65	0.03	0.65	0.03	0.65	0.02	0.65	0.02
Pos4	0.77	0.03	0.77	0.03	0.75	0.03	0.75	0.03
Neg1	0.64	0.03	0.64	0.03	0.64	0.03	0.64	0.03
Neg2	0.88	0.03	0.88	0.03	0.88	0.03	0.88	0.03
Neg3	0.15	0.00	0.15	0.00	0.15	0.00	0.15	0.00

Note. The positively worded (Pos1–Pos4) and negatively worded (Neg1–Neg3) indicators had a four-category

Likert response scale (1 = *agree a lot*, 2 = *agree a little*, 3 = *disagree a little*, 4 = *disagree a lot*). The table reports unstandardized parameter estimates and standard errors of model estimations in which sampling weights were applied. Fixed latent means and variances are depicted without standard errors. The reported class sizes are the final class proportions based on the estimated model in percent. Cases with missing values on all variables were excluded from the analyses. The effective sample size amounted to $n = 5,863$ students.

Appendix 5.6

Model Estimation Results of the Factor Mixture Analysis with Invariant Negatively Worded Items (Neg. Invariant FMA), and the Non-Invariant Factor Mixture Analysis (Non-Invariant FMA) for the Reading Self-Concept Scale in PISA in the USA

	Neg. invariant FMA				Non-invariant FMA			
	C#1 92%		C#2 8%		C#1 91%		C#2 9%	
	Par.	SE	Par.	SE	Par.	SE	Par.	SE
<i>Factors</i>								
Mean α	0		0.72	0.10	0		0	
Variance ϕ	1		1.74	0.17	1		1	
<i>Loadings λ</i>								
Pos1	-0.59	0.01	-0.59	0.01	-0.58	0.02	0.76	0.12
Pos2	-0.56	0.01	-0.56	0.01	-0.58	0.01	0.67	0.06
Pos3	-0.55	0.01	-0.55	0.01	-0.54	0.01	0.71	0.06
Neg1	0.60	0.01	-0.55	0.04	0.59	0.02	0.69	0.05
Neg2	0.48	0.01	-0.53	0.03	0.49	0.01	0.68	0.04
Neg3	0.49	0.01	-0.50	0.03	0.50	0.01	0.63	0.05
<i>Intercepts ν</i>								
Pos1	3.14	0.01	3.14	0.01	3.15	0.02	2.62	0.21
Pos2	2.94	0.01	2.94	0.01	2.92	0.02	2.75	0.07
Pos3	3.11	0.01	3.11	0.01	3.13	0.02	2.55	0.16
Neg1	1.85	0.02	3.00	0.07	1.84	0.02	2.65	0.12
Neg2	2.41	0.02	3.13	0.06	2.41	0.02	2.69	0.08
Neg3	2.06	0.02	2.90	0.07	2.07	0.02	2.48	0.11
<i>Res. var. Ψ</i>								
Pos1	0.16	0.01	0.16	0.01	0.17	0.02	0.17	0.02
Pos2	0.24	0.01	0.24	0.01	0.23	0.01	0.23	0.01
Pos3	0.21	0.01	0.21	0.01	0.20	0.02	0.20	0.02
Neg1	0.29	0.01	0.29	0.01	0.28	0.01	0.28	0.01
Neg2	0.46	0.01	0.46	0.01	0.46	0.01	0.46	0.01
Neg3	0.36	0.01	0.36	0.01	0.36	0.01	0.36	0.01

Note. The positively worded (Pos1–Pos3) and negatively worded (Neg1–Neg3) indicators had a four-category

Likert response scale (1 = *strongly disagree*, 2 = *disagree*, 3 = *agree*, 4 = *strongly agree*). The table reports unstandardized parameter estimates and standard errors of model estimations in which sampling weights were applied. Fixed latent means and variances are depicted without standard errors. The reported class sizes are the final class proportions based on the estimated model in percent. Cases with missing values on all variables were excluded from the analyses. The effective sample size amounted to $n = 4,703$ students.