Comparing the Roles of Cognitive Abilities and Personality Traits in Inconsistent Responding: The Case of a Mixed-worded Self-esteem Scale in the German National Educational Panel Study

Jianan Chen

UNIVERSITY OF OSLO

Master of Science in Assessment, Measurement and Evaluation

30 credits master thesis

CEMO: Centre for Educational Measurement
Faculty of Educational Sciences

Spring 2023

Popular Abstract

Mixed-worded scales which contain both positively worded items such as "I have some positive attitude towards myself" and negatively worded items such as "Sometimes I really feel useless" are often used in surveys to keep respondents paying attention. However, some respondents tend to give inconsistent responses, that is, agreeing or disagreeing with both positively and negatively worded items, which leads to meaningless response data. Recent research suggests that such inconsistent responses are more likely among students with lower cognitive abilities or certain personality traits. This study examines both explanations at the same time. A sample of n = 4,938 Grade 5 students in Germany was classified into an inconsistent (11%) and a consistent (89%) group given their responses on a self-esteem scale, and the students' group memberships were further related to their cognitive abilities and personality traits. The results suggested that both having lower cognitive abilities (especially a lower reading comprehension ability) and some personality traits (especially being less conscientious) were related to inconsistent responding, while ability played a more important role. The use of mixed-worded scales requires more caution among young learners with lower reading abilities under low-stakes contexts.

Acknowledgements

The journey of writing this thesis has been a challenging yet rewarding experience, like finding a destination step by step to find a destination rather than simply driving there. Along the way, I was fortunate to collect treasures that will undoubtedly benefit me in my future pursuits.

Many thanks to my supervisors, Johan and Isa, for their inspiration for the topic, guidance in methodology and writing, and constructive feedback. Their mentorship and support have been invaluable in shaping this work, and I could not have accomplished it on time without their help. I would like to thank my partner, Haifeng, for his constant thoughtfulness and encouragement. I am also grateful to my parents for their unconditional trust and to my colleagues at CEMO for their insights and support.

Finally, I would like to acknowledge the National Educational Panel Study (NEPS) for providing the German data used in this study: Starting Cohort 3 – 5th Grade, doi:10.5157/NEPS:SC3:5.0.0. From 2008 to 2013, NEPS data were collected as part of the Framework Programme for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research (BMBF). As of 2014, the NEPS survey is carried out by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg in cooperation with a nationwide network.

Comparing the Roles of Cognitive Abilities and Personality Traits in Inconsistent Responding: The Case of a Mixed-worded Self-esteem Scale in the German National Educational Panel Study

Mixed-worded survey scales are commonly used to ensure respondent attentiveness. Respondents need to switch response scales according to the wording direction (i.e., positive or negative wording) when answering an item. However, some respondents tend to deliver inconsistent responses (i.e., agreeing or disagreeing with both positively and negatively worded items) on such mixed-worded items in practice, which poses a validity concern. Recent research has proposed two potential directions of individual determinants driving inconsistent responding: cognitive abilities and personality traits. Due to low cognitive abilities, some respondents might be less effective in providing consistent responses. As for personality traits, some respondents might be careless due to low conscientiousness. The study contributes to simultaneously investigating the roles of both in inconsistent responding from an individual-centered perspective. Adopting a factor mixture analysis model, inconsistent responding was investigated as a function of individuals' cognitive ability and personality, among $n = 4,938$ Grade 5 students from the German National Educational Panel Study (NEPS). 11% of the students were classified as inconsistent responders on Rosenberg's self-esteem scale, and their class memberships were further related to four cognitive abilities (cognitive reasoning, cognitive speed, reading comprehension, reading speed) and five personality traits (conscientiousness, neuroticism, extraversion, agreeableness, openness). In general, the model comparison indicated that both ability and personality predictors matter with a more prominent role for ability (especially reading comprehension). The implications for the use of mixed-worded scales and future research directions are discussed.

Keywords: inconsistent responding, mixed-worded scale, cognitive abilities, Big Five personality traits, factor mixture analysis, NEPS

A mixed-worded scale is a commonly used survey design, which contains items with both positive and negative (or regular and reverse-keyed) wordings of the construct intended to measure. The opposite wording can be generated by either using negations (e.g., "no", "not", "un-", "non-") or antonyms (e.g., "happy" vs. "sad") (Menold, 2020). For instance, Rosenberg's (1965) self-esteem scale includes positively worded (PW) items such as "I feel that I have a number of good qualities", and negatively worded (NW) items such as "At times I think I am no good at all" (negation) and "I certainly feel useless at times" (antonyms). Respondents are expected to switch the response scale according to the wording direction of the item on a mixed-worded scale. If the scale measures the same construct, a response towards the upper side of the response scale on a PW item would imply a response towards the lower side of the response scale on an NW item, and vice versa. Take the self-esteem scale for example, a consistent responder is expected to agree on PW items and disagree on NW items on the self-esteem scale to express a high self-esteem level.

Mixed-worded scales are often employed in surveys as a type of quality assurance, as the respondents are required to attentively read each item to give consistent responses (Steinmann, Sánchez, et al., 2022). Mixed wording is also used to reduce systematic response pattern biases caused by acquiescence behaviors, which refers to the tendency that respondents are willing to agree rather than disagree on an item regardless of its content (Paulhus, 1991; Podsakoff et al., 2003). In addition, including mixed-worded items in surveys enables a post-hoc check of acquiescence level in the data (Buchholz, 2022). Mixed-worded scales have been widely adopted in the background questionnaires of large-scale assessments in education, such as the Progress in International Reading Literacy Study (PIRLS), the Trends in International Mathematics and Science Study (TIMSS), the Programme for International Student Assessment (PISA), and the German National Educational Panel Study (NEPS), for investigating, for instance, student's attitudes and beliefs. However, previous studies indicated that mixed-worded scales in survey assessments should be used with caution since they could lead to unintended consequences, threatening the validity of survey data and conclusions drawn for policy recommendations (e.g., Marsh, 1996; Steedle et al., 2019; Steinmann, Sánchez, et al., 2022). Some of these studies take an instrument-centered

perspective using factor analysis or item response theory, and others take an individual-centered perspective focusing on individuals with differential response patterns.

Wording Effect from an Instrument-centered Perspective

Previous empirical studies found that the intercorrelations across PW and NW items are often attenuated, in other words, PW and NW items are less negatively correlated than expected on a unidimensional mixed-worded scale (Dunbar et al., 2000; Marsh, 1986; Steinmann, Strietholt, et al., 2022). Therefore, the use of mixed-worded scales could risk lower reliability (e.g., Barnette, 2000) and poor model fit of a unidimensional factor model (e.g., Marsh, 1996). Studies also showed that introducing method factors or correlated uniqueness for PW items and/or NW items significantly improved model fit (e.g., DiStefano & Motl, 2009; Wang et al., 2015). The common finding is that mixed-worded scales lead to more complex latent structures than intended. This can be interpreted as a systematic difference in responses affected by the wording direction of the item, which is often called the item wording effect, method effect, or keying effect. Traditionally, method effects have been regarded as construct irrelevant variance and should be minimized (Marsh, 1996).

Some of the studies from an instrument-centered perspective have also found a relation between the "method factors" and respondent characteristics, such as cognitive abilities (e.g., Dunbar et al., 2000; Gnambs & Schroeders, 2020; Marsh, 1986; Michaelides, 2019), and self-reported conscientiousness and neuroticism/emotional stability (e.g., Michaelides, Koutsogiorgi, et al., 2016; Michaelides, Zenger, et al., 2016; Quilty et al., 2006). This implies that population heterogeneity may exist in responses to mixed-worded scales instead of the "method factors" equally affecting all respondents. Hence, an alternative, individual-centered perspective might be relevant.

Inconsistent Responding from an Individual-centered Perspective

Another set of studies has focused on wording effects from an individual-centered perspective, with the aim of detecting consistent and inconsistent respondents (e.g., Kam & Chan, 2018; Steedle et al., 2019). As mentioned before, a consistent responder will switch the response scale according to the item wording direction. Yet in practice, there are some

responders who do not switch the response scale accordingly. A typical inconsistent responder may strongly agree to both PW and NW items, with a resulting lack of internal consistency in their responses, given that the items measure the same construct.

A small proportion of the respondents were found to deliver inconsistent responses (sometimes also referred to as misresponse or inattentive/careless/insufficient effort responding) on mixed-worded scales. For instance, Steedle et al. (2019) identified around 10% of the participants as inconsistent responders in a social-emotional learning assessment in the US. Steinmann, Sánchez, et al. (2022) identified 2%-36% of students as inconsistent responders across education systems and across three mix-worded self-concept scales in the joint PIRLS/TIMSS 2011 assessment. Steinmann, Strietholt, et al. (2022) identified that 7%-20% of the participants showed inconsistent responses on four student questionnaire scales across five datasets from Germany, Australia, and the US.

The mechanism behind inconsistent responding behavior has not been clearly examined and it could theoretically arise at four different cognitive stages of responding (Weijters & Baumgartner, 2012), namely "(1) comprehension (attending to a question and interpreting it), (2) retrieval (generating a retrieval strategy and then retrieving relevant beliefs from memory), (3) judgment (integrating the information into a judgment), and (4) response (mapping the judgment onto the response categories provided and answering the question)". However, recent research has revealed two directions of individual determinants when it comes to explaining inconsistent responding. One suggests that it may be associated with lower cognitive abilities, including reading abilities. The other proposes that inconsistent responding may be linked to personality traits, such as conscientiousness and neuroticism/emotional stability.

Cognitive Abilities and Inconsistent Responding

It is plausible that mixed-worded items could be challenging for some respondents to interpret, due to the difficulty of processing NW items. The conjecture is that individuals with higher cognitive abilities may be more intelligent and quick in processing, hence it could be less resource intensive for them to deal with the mixed-worded scales which leads to fewer

mistakes or less inconsistency, and individuals with higher reading abilities may be better at understanding the meaning of the questions and hence more likely to respond in a consistent manner. In contrast, individuals with lower cognitive and reading abilities may struggle to pick up the wording differences and fail to give consistent responses. Moreover, previous empirical evidence has supported that lower reading ability, cognitive reasoning, and academic competence (e.g., high school grade point average) are correlated with a higher risk of inconsistent responding (e.g., Bolt et al., 2020; Marsh, 1986; Steedle et al., 2019; Steinmann, Strietholt, et al., 2022). This also implies that inconsistent responding can be expected to happen more likely in young kids such as students in primary schools than teenagers such as students in secondary schools since the former's reading ability has not yet fully matured (e.g., Steinmann, Strietholt, et al., 2022).

Personality Traits and Inconsistent Responding

Previous studies focused on correlations between personality traits and wording effects related to negative item wording, mainly from an instrument-centered perspective. Careless responding is often suspected to be a source of wording effect (see e.g., Schmitt & Stuits, 1985). Quilty et al. (2006) found that conscientiousness and emotional stability (i.e., the opposite of neuroticism) are positively related to the negative item wording effect, on the self-esteem scale. Michaelides, Zenger, et al. (2016) identified emotional stability as the most significant personality trait in relation to mixed-wording effects, with a negative relation to the positive wording effect and a positive relation to the negative wording effect, on the self-esteem scale. These findings indicate that individuals who are less conscientious and more neurotic are more likely to endorse NW items, and the more neurotic individuals may endorse PW items more strongly as well.

Although the correlation between personality and method effects from an instrument-centered perspective can hardly be directly interpreted in the context of individual response inconsistency, it suggests that individual differences in some personality traits could affect the perception of PW and NW items and contribute to explaining differential response patterns and inconsistent responses. It is plausible to expect a negative relation between

conscientiousness and inconsistent responding because some respondents may just not be careful enough to notice the change in item wording direction and hence not be able to shift the response scale accordingly.

The Roles of Cognitive Abilities and Personality Traits in Inconsistent Responding

Few previous studies have compared the roles of cognitive abilities and personality traits in inconsistent responses. However, an empirical study suggested that the cause of misresponse on mixed-worded scales may be more influenced by difficulty of the questionnaire items, rather than inattention (Baumgartner et al., 2018). In this study, they proposed two mechanisms (difficulty vs. inattention, i.e., lack of ability vs. lack of motivation) to explain misresponses to three types of items: negated items (e.g., talkative vs. not talkative), polar opposite items (items with an opposite core concept to the regular items; e.g., talkative vs. quiet), and reversed items (e.g., talkative vs. not talkative; talkative vs. quiet). By conducting a factor analysis with eye-tracking data, they found that the three types of items varied in gaze durations and degrees of misresponse. Negated items had longer gaze duration (i.e., received greater attention) than nonnegated items and it helped to prevent misresponse; Polar opposite items also received greater attention than regular items but still resulted in higher misresponse; Reversed items were not processed significantly more than non-reversed items, but it did not have a negative impact on response consistency. The evidence implies that paying greater attention to NW items does not always guarantee the avoidance of inconsistent responses, whilst not giving extra attention to NW items does not necessarily lead to inconsistent responses either. Contrary to popular belief, inattention may not be the primary reason behind inconsistent responding. Instead, the difficulty level of the items, which can challenge the cognitive and reading abilities of respondents, may play a more significant role.

In addition, Steinmann, Strietholt, et al. (2022)'s study found a significant correlation between inconsistent responding and reading ability but not with conscientiousness among Grade 9 students. These findings imply that the lack of reading and cognitive abilities is probably a more crucial factor than conscientiousness (or other potential personality traits) in driving inconsistent responses.

The present study

The current study investigates whether individual differences in ability and personality are associated with being classified as inconsistent responders on a mixed-worded self-esteem scale, using data from the German National Educational Panel Study (NEPS). The classification of in/consistent responders at the individual level is based on the constrained factor mixture model introduced by Steinmann, Strietholt, et al. (2022). Specifically, there are three research questions:

1. To what extent are students' cognitive abilities related to (being classified as) inconsistent responding? 2. To what extent are students' personality traits related to (being classified as) inconsistent responding? 3. Comparing ability and personality, which one contributes more in explaining inconsistent responding?

The first expectation is that students with higher cognitive abilities would be less likely to be classified as inconsistent responders. The second expectation is that students who are more conscientious would be less likely to be classified as inconsistent responders, and the study further explores personality traits including neuroticism, extraversion, agreeableness, and openness. The third expectation is that ability plays a more important role than personality in explaining individuals' response inconsistency, as discussed earlier.

Note that this study is partly a replication of Steinmann, Strietholt, et al. (2022)'s study, which investigated and found a negative association between reading comprehension ability and inconsistent responding. One contribution of the present study is to further include other cognitive abilities and personality traits and compare the relative roles of the two. Upon searching the recent literature on inconsistent responding, there appears to be a scarcity of studies investigating the association between personality traits and inconsistent responding from an individual-centered perspective or studies simultaneously examining relevant ability and personality factors in individual response inconsistency. Hence, the present study makes a unique contribution to filling the research gap, understanding the relative roles of ability and personality in driving individual inconsistent responses, and providing implications for the use of mixed-worded scales and the validity of survey data.

Method

The National Educational Panel Study (NEPS) is a large-scale educational study in Germany led by Leibniz Institute for Educational Trajectories (LIfBi). It provides longitudinal data on individual educational processes and outcomes throughout life, covering all stages from birth and early childhood education to adult education and lifelong learning (Blossfeld & Roßbach, 2019). NEPS data assesses a mixed-worded scale, cognitive abilities as well as personality traits, which are three important elements of this study.

Sample

The current study used the NEPS data of the Starting Cohort 3 (SC3), which targets Grade 5 students in Germany and was initiated during the 2010/2011 academic year. The original sample of SC3 in the NEPS study excluded students in vocational schools or schools with predominantly foreign teaching languages, as well as students unable to comply with normal testing procedures in regular schools (Blossfeld & Roßbach, 2019). The sample followed a two-stage stratified cluster sampling procedure with schools as the primary sampling units in the first stage and two classes randomly selected (for regular schools) or a full sample (for special schools) of Grade 5 in the second stage. For more details of the sampling procedure such as explicit and implicit stratification, see the NEPS technical report for weighting (Steinhauer & Zinn, 2016).

Our study included all the students who participated in the first wave (i.e., the original sample), with the exception of those from special needs schools or those who were part of the oversampling for migrant students. The additional migrant sample was excluded because its survey and test instruments differ from the main sample. Additionally, 34 students were excluded from the study as they did not provide any responses (i.e., with all ten items missing) on the self-esteem scale, and had higher agreeableness and reading speed compared with the sample. As a result, the effective sample in this study contained $n = 4,972 - 34 = 4,938$ Grade 5 students (age in years $M = 11.04$, $SD = .64$, gender approximately balanced 50/50) from 203 schools.

In addition to data from Wave 1, the self-reported Big Five personality data from Wave

3 was also used because the students' personality traits were not measured in the first two waves. Among the 4,938 students in Wave 1, 684 (14%) students did not participate in Wave 3. As a consequence, the proportions of missing values on Wave 1 variables (i.e., the mixed-worded items and the cognitive abilities) are relatively low (<5%), while the missingness rate is on average 18% for the Big Five personality trait scores from Wave 3 (see Table 4 in the result section). The attrition in Wave 3 data can be due to different reasons, e.g., grade repetition, moving from a higher-track school to a lower-track school, or non-selective school switching (some schools provide a 6-year primary education), according to the specific education system of each federal state of Germany. The missing values in the effective sample were handled with a multiple imputation approach, which will be described in detail later on.

Measures

The measures used in this study were surveyed or tested in the German language and administered in a paper-pencil mode. Specifically, the students' responses to the mixed-worded self-esteem scale and their performance in four reading- and cognitive-related ability tests were extracted from Wave 1 and their self-reported five personality traits were extracted from Wave 3 (see Table 1 for a summary).

Table 1

Summary of NEPS Data Used in the Current Study

	Variable	Description	Wave
	Mixed-worded Self-esteem	Ten items; five-point Likert scale	1
Ability	Reading Comprehension	WLE score with mean of zero	1
	Reading Speed	sumscore of correct items (51 items in total)	1
	Cognitive Reasoning	sumscore of correct items (12 items in total)	1
	Cognitive Speed	sumscore of correct items (93 items in total)	1
Personality	Conscientiousness	Two items; five-point Likert scale	3
	Extraversion	Two items; five-point Likert scale	3
	Neuroticism	Two items; five-point Likert scale	3
	Agreeableness	Three items; five-point Likert scale	3
	Openness	Two items; five-point Likert scale	3

Note. WLE score = weighted maximum likelihood estimate score

Mixed-worded Scale: Self-esteem

The student survey included a German version of the Rosenberg Self-esteem Scale, on which the classification of inconsistent respondents in this study was conducted. The scale was developed to measure an individual's favorable or unfavorable self-perception (Rosenberg, 1965). It contained ten items (variables: 't66003a'-'t66003j'), five of which are positively worded and the remaining five are negatively worded (see Table 2 for a translated version in English; the German version is presented in Table C5). The response scale is a five-point Likert scale, ranging from 1 (Does not apply at all) to 5 (Applies completely).

Table 2

Item Wording of the Self-esteem Scale in NEPS Starting Cohort 3, Wave 1 (Grade 5)

Item	To what extent do the following statements apply to you?
PW1	All in all, I am satisfied with myself.
NW1	Now and then I think that I'm not good for anything.
PW2	I have some positive attributes.
PW3	I can do many things just as well as most other people.
NW2	I am afraid there is not much I can be proud of.
NW3	Sometimes I really feel useless.
PW4	I consider myself a valuable person, at least I am not less valuable than the others.
NW4	I wish I could have more respect for myself.
NW5	All in all, I tend to consider myself a loser.
PW5	I have a positive attitude towards myself.

Note. PW represents positively worded items; NW represents negatively worded items. Response scale: Does not apply at all = 1; Does rather not apply = 2; Partly = 3; Does rather apply = 4; Applies completely = 5. The items are ordered in the original sequence as presented in the questionnaire. The original student questionnaire was distributed in German.

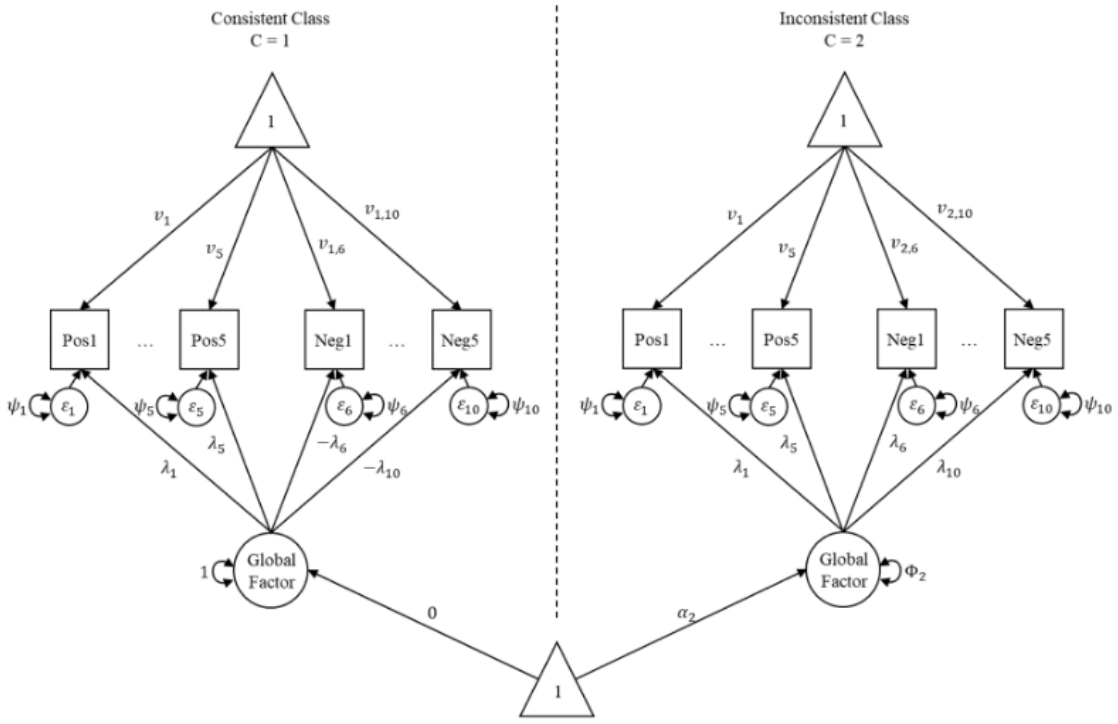
Outcome Variable: Classification as Inconsistent Responder

Factor Mixture Model. To classify a student as a consistent or an inconsistent responder on the self-esteem scale, the constrained factor mixture analysis (FMA) model proposed by Steinmann, Strietholt, et al. (2022) was adopted (see Figure 1). The model assumed the existence of two latent classes within the target population, namely a consistent class and an inconsistent class. Adding class-invariant constraints for the PW items and

reflected loading constraints for the NW items allows for the identification of the two classes with different response patterns. The core feature of the model is that the NW items were assumed to have opposite factor loadings across the two classes (i.e., $\lambda^{-}_{1,i} = -\lambda^{-}_{2,i}$), and the PW items were assumed to have the same intercepts and factor loadings across the two classes (i.e., $v^{+}_{1,i} = v^{+}_{2,i}$ and $\lambda^{+}_{1,i} = \lambda^{+}_{2,i}$). The factor loadings of the PW items were constrained to be positive (i.e., $\lambda^{+}_{k,i} > 0$) and the factor loadings of the NW items were constrained to be negative (i.e., $\lambda^{-}_{k,i} < 0$) in the consistent class.

Figure 1

Constrained Factor Mixture Analysis Model to Classify Inconsistent Responders



Note. This model followed standard path diagram conventions and includes five positively worded (Pos1–Pos5) and five negatively worded items (Neg1–Neg5). Reprinted under the terms of CC-BY-NC from "A constrained factor mixture analysis model for consistent and inconsistent respondents to mixed-worded scales." by I. Steinmann, R. Strietholt and J. Braeken, 2022, *Psychological Methods*.

Classification. The outcome variable of interest, the binary latent class of consistent or inconsistent responder, was estimated based on the students' observed item response patterns

on the self-esteem scale. The FMA model was estimated and the students were classified into two groups based on their maximum posterior class membership probability. Average class membership probabilities and entropy were used to evaluate classification precision (Masyn, 2013).

Main Predictors: Reading and Cognitive Abilities

As part of the measurement of individual competencies and skills in Wave 1, NEPS administered two non-verbal cognitive ability tests and two more reading-specific ability tests.

Cognitive Reasoning. The NEPS reasoning test (NEPS-MAT) is structured as a Raven's progressive matrices test (Raven, 1941) measuring non-verbal reasoning, which is a key indicator of fluid intelligence (Gottfredson, 1997). The test consisted of three sets of four items each (i.e., 12 items in total), with a time limit of three minutes per set (i.e., 9 minutes in total). The students needed to figure out the pattern logic of the geometrical elements presented in each item and select the correct solution for a missing field from the given options (Haberkorn & Pohl, 2013). The sumscore correct (with a maximum of 12 points) was recorded as variable 'dgg5_sc3b'.

Cognitive Speed. The NEPS Picture Symbol Test (NEPS-BZT) measured perceptual speed, reflecting the speed of information processing. The students needed to match figures or numbers with graphical symbols as quickly as possible, by entering the correct figures/numbers for the presented symbols in line with the given answer key. This is based on an enhanced version of the Wechsler family's Digit-Symbol Test for assessing intelligence developed by Lang et al. (2007) but requires performance in the opposite direction. The test consisted of three sets of 31 items (i.e., 93 items in total), with a time limit of 30 seconds per set (i.e., 90 seconds in total). The sumscore correct (with a maximum of 93 points) was recorded as variable 'dgg5_sc3a'.

Reading Comprehension. The reading comprehension test addressed "competent handling of written texts in different and typical everyday situations", with item formats of multiple-choice, decision-making tasks and matching tasks; it covered five text functions (informational, commenting or argumenting, literary, instructional, and advertising). Within a time span of 28 minutes, the students needed to complete the test which consisted of five texts

corresponding to the five text functions and five to seven items for each text (Gehrer et al., 2012). For reading comprehension, a weighted maximum likelihood estimate (WLE) score (variable: 'reg5_sc1') was used. The WLE score was scaled to have a mean of zero (Pohl & Carstensen, 2012). Thus, a positive WLE score indicates an above-average reading comprehension ability and a negative WLE score indicates a below-average reading comprehension ability. The reading comprehension score was corrected for the test position since there were two booklets in which reading comprehension was tested either before or after a mathematical competency test. It should be noted that the correction did not apply to other covariates tested or investigated at a fixed position in the competency test or questionnaire for all the students.

Reading speed. The reading speed test aimed to assess the respondents' automatized reading processes; it had 51 items and required the respondents to rate the short sentences as either true or false, using common knowledge, within exactly two minutes (Zimmermann et al., 2012). The reading speed score (variable: 'rsg5_sc3') was given by the number of correctly-judged sentences during the time limit.

Main Predictors: Five Personality Traits

In wave 3 (i.e., grade 7), NEPS introduced the Big Five self-reported personality measures in the student survey. The scale contained 11 items measuring five personality traits (neuroticism, conscientiousness, extraversion, agreeableness, and openness). See Table 3 for a translated version in English from NEPS; the German version is presented in Table C6). The response scale was a five-point Likert scale, ranging from 1 (Does not apply at all) to 5 (Applies completely). Generated average scores of five traits (variables: 't66800a_g1'-'t66800e_g1') were used as personality measures. A higher trait score indicates that the personality trait applies more to the person. Although measured two years after other variables, students' personality traits were considered relatively stable over time (Borghuis et al., 2017; John & Srivastava, 1999).

Table 3

Item Wording of the Big-Five Scale in NEPS Starting Cohort 3, Wave 3 (Grade 7)

To what extent do the following statements apply to you?	Trait
a) I am quite cautious, reserved.	Extraversion
b) I trust other people easily, I believe in the goodness in people.	Agreeableness
c) I am easy-going and tend to be a bit lazy.	Conscientiousness
d) I am relaxed and don't get easily stressed.	Neuroticism
e) I do not care much about arts.	Openness
f) I am out-going and sociable.	Extraversion
g) I tend to be critical of other people.	Agreeableness
h) I am thorough.	Conscientiousness
i) I easily get nervous and self-conscious.	Neuroticism
j) I have an active imagination, I am an imaginative person.	Openness
k) I am considerate, sensitive.	Agreeableness

Note. Response scale: Does not apply at all = 1; Does not really apply = 2; Partially applies = 3; Applies to some extent = 4; Applies completely = 5. The items are ordered in the original sequence as presented in the questionnaire. The original student questionnaire was distributed in German.

Statistical Analysis

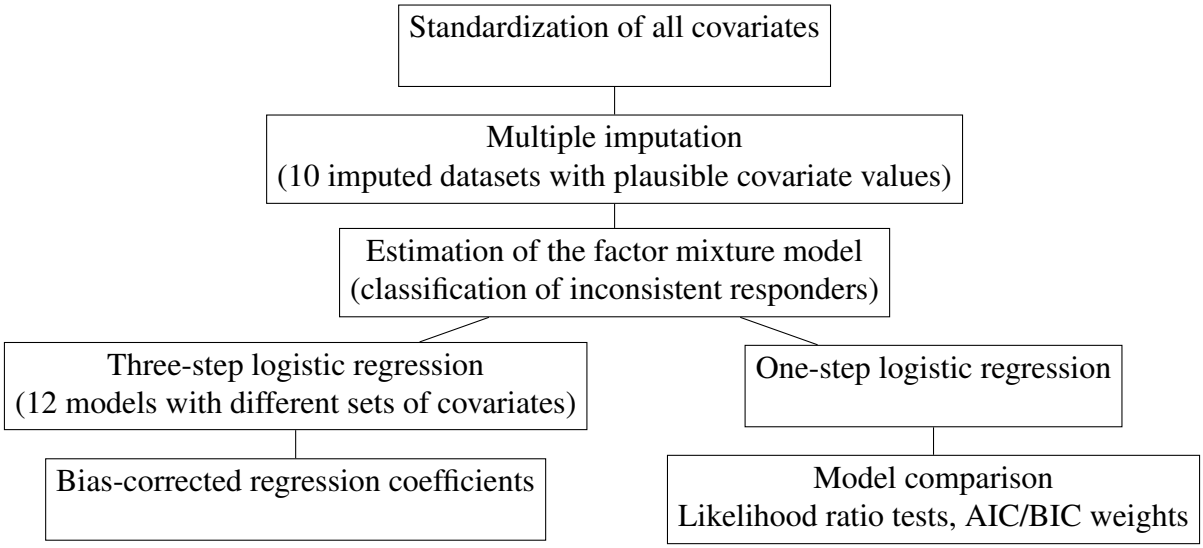
All statistical analyses were run through a combination of the statistical software environments R Version 4.2.1(R Core Team, 2020) for pre- and post-processing of results and Mplus Version 8.3 (Muthén & Muthén, 1998-2017) for model estimation. This study's analysis steps are summarized in Figure 2.

All covariates (i.e., ability and personality measures) were standardized prior to further analyses and a multiple imputation approach (Enders, 2010) was adopted to deal with missing data on the covariates. A fully saturated model including all self-esteem items as auxiliary variables together with the main predictors described above was used to generate 10 imputed datasets containing plausible values for missing data on all covariates through Markov Chain Monte Carlo simulation (Asparouhov & Muthén, 2022). Note that for the imputation model, missing responses to self-esteem items were not imputed for the consistency of classification, and personality items were treated as categorical variables to remain as close to the data as possible.

The factor mixture model by Steinmann, Strietholt, et al. (2022) was estimated for the

Figure 2

Diagram summarizing Analysis Steps of the Study



self-esteem scale, treating item responses as interval measures, using full information maximum likelihood in Mplus (5000 random sets of starting values for the initial estimation stage and 500 optimizations for the final stage).

Logistic regression models related the estimated class membership of the students to their abilities and personality traits for each of the 10 imputed datasets. A three-step estimation (Asparouhov & Muthén, 2014) was used to account for classification errors in the estimated class membership. Following a model comparison strategy, 12 models including different sets of covariates (single predictors, ability covariate block of predictors, personality covariate block of predictors, and the combination of ability and personality) were run. The model with all the predictors is of main interest to the present study. For each model, the results were combined across imputations following Rubin's rules (Rubin, 1978).

Given that the three-step procedure as implemented in Mplus does not provide model comparison measures for the logistic regression model, these were extracted from the unadjusted logistic regression model with as an outcome the maximum posterior membership classification. Likelihood ratio tests for nested model comparison, and Akaike information criterion (AIC) and Bayesian information criterion (BIC) weights were reported (Wagenmakers & Farrell, 2004) for the full sets of models.

For all analyses reported here, the NEPS cross-sectional Wave 1 student weights were

used to account for non-response and unequal selection probability during sampling, and robust huber-white sandwich errors were used to account for students being nested in schools.

Results

Descriptives

Among self-esteem items, PW items had an average mean of around 4 which corresponded to a "does rather apply" response, while non-reverse-coded NW items had an average mean slightly above 2 which corresponds to a "does rather not apply" response (Table 4). PW items showed negative skewness while NW items showed positive skewness. Reading comprehension had an average WLE score close to zero, and means of other abilities represented their average sum scores. In terms of personalities, an average mean of around 3 corresponded to a response of "partially applies". Self-esteem items and ability-related variables had missing rates of up to 5%, while 17% to 18% of the personality-related variables were missing in the original effective sample. As mentioned earlier, the high missing rate for personality traits is in large part due to the attrition in Wave 3 data. Note that the ability and personality variables were standardized in the later, main analyses.

Table 5 presents the correlations between the predictors. Between abilities, the correlations were positive, ranging from 0.10 to 0.44. Across personalities and abilities, the correlations were close to zero. Between personalities, the correlations were mostly zero with some exceptions of low positive or negative correlations of around 0.20.

Inconsistent Responders: Factor Mixture Model and Classification

Table 6 summarizes the main results of the factor mixture model for identifying inconsistent responders. The factor mixture model estimated that the probability of belonging to the inconsistent class was 14%, which was much smaller than the probability of belonging to the consistent class (86%) in the sample. The estimation results were in line with the results reported in (Steinmann, Strietholt, et al., 2022). The unstandardized factor loadings of PW items were constrained to be positive and identical across the consistent and inconsistent classes, while those of NW items were constrained to be opposite across the two classes

Table 4

Descriptive Statistics of variables

	Variable	Mean	SD	Skewness	Excess Kurtosis	Missing (%)
Self-Esteem Items	PW1	4.28	0.86	-1.09	0.85	1%
	NW1	2.45	1.31	0.44	-0.98	3%
	PW2	4.08	0.90	-0.78	0.21	2%
	PW3	4.01	0.96	-0.79	0.15	2%
	NW2	2.19	1.31	0.85	-0.46	3%
	NW3	1.84	1.17	1.32	0.75	2%
	PW4	3.89	1.24	-0.94	-0.17	3%
	NW4	2.74	1.37	0.18	-1.16	5%
	NW5	1.79	1.15	1.40	0.98	4%
	PW5	4.03	1.06	-0.98	0.36	3%
Ability	Reading Comprehension	-0.06	1.25	0.18	0.30	0%
	Reading Speed	20.89	6.78	1.15	3.92	0%
	Cognitive Reasoning	6.83	2.60	-0.37	-0.38	0%
	Cognitive Speed	44.02	13.45	0.73	1.57	0%
Personality	Conscientiousness	3.23	0.86	-0.05	-0.24	17%
	Extraversion	3.42	0.78	0.01	-0.17	18%
	Neuroticism	2.83	0.82	0.09	-0.11	17%
	Agreeableness	3.45	0.65	-0.31	0.50	18%
	Openness	3.47	0.94	-0.21	-0.49	17%

Note. All the descriptive statistics are weighted. PW represents positively worded items, and NW represents negatively worded items. Sample size n = 4, 938.

(Table 6). The congeneric reliability (i.e., coefficient omega) for the consistent class was estimated to be 0.78. The positive correlations between PW and NW items implied that if they score high/low on any item, they do this as well on the other items, regardless of wording (note that the NW items were reverse-coded for computing and interpreting the reliability of the mixed-worded scale).

The average intercept of PW and NW items in the consistent class were 4.12 and 2.02, respectively. In contrast, in the inconsistent class, the average intercept of PW and NW items were 4.12 and 3.82, respectively. These findings indicated that a consistent responder with average self-esteem was expected to rate about 2.1 scale points lower on NW items than on PW items, while an inconsistent responder was expected to give similar average scores on PW and non-reverse-coded NW items.

Based on their most likely latent class membership, 568 (11%) of the students were

Table 5

Correlations of the Ability-related and Personality-related Predictors

	Ability	Reading Comprehension	Reading Speed	Cognitive Reasoning	
Ability	Reading Speed	0.34			
	Cognitive Reasoning	0.44	0.20		
	Cognitive Speed	0.10	0.30	0.15	
	Ability	Reading Comprehension	Reading Speed	Cognitive Reasoning	Cognitive Speed
Personality	Conscientiousness	0.01	0.01	-0.03	0.00
	Extraversion	0.04	0.07	-0.01	0.03
	Neuroticism	-0.05	-0.05	-0.05	-0.03
	Agreeableness	-0.04	-0.03	0.01	0.03
	Openness	0.10	0.04	0.06	0.04
	Personality	Conscien-tiousness	Extra-version	Neurot-icism	Agree-ableness
Personality	Extraversion	0.03			
	Neuroticism	-0.05	-0.23		
	Agreeableness	0.28	0.00	-0.05	
	Openness	0.09	0.09	0.02	0.20

Note. All the correlations are weighted. Sample size n = 4, 938.

assigned to the inconsistent class and the others were assigned to the consistent class, after weighting. The average inconsistent and consistent latent class probabilities for most likely inconsistent class membership were 0.87 and 0.13, respectively. The average inconsistent and consistent latent class probabilities for most likely consistent class membership were 0.04 and 0.96, respectively. The entropy value was 0.831, indicating an appropriate classification quality.

Additionally, the model-implied item intercorrelations in both classes were compared in Figure 3, and the positive correlations between PW and non-reverse-coded NW items were shown in the inconsistent class. Combining the evidence, it implied that the inconsistent class was indeed "inconsistent" by answering the items in a similar way, no matter the directions of the item wording.

Inconsistent Responder classification as a function of Ability and Personality

The results of the three-step logistic analyses (Table 7) together with the one-step logistic model comparison (Table 8) showed that both ability-related and personality-related

Table 6

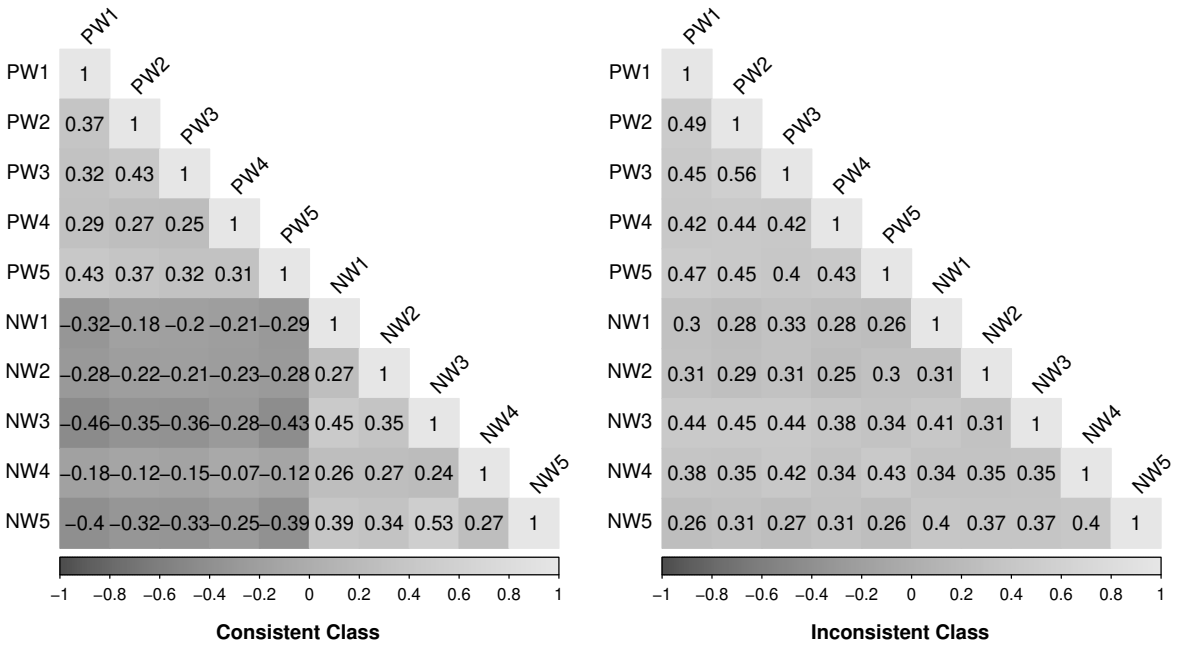
Constrained Factor Mixture Analysis for the Self-esteem Scale in NEPS SC3 Grade 5 students

	Parameters	Consistent Class (86%)		Inconsistent Class (14%)	
		Unstandardized	Standardized	Unstandardized	Standardized
Factor	Mean	0	0	-0.81 (0.56)	-0.63
	Variance	1	1	1.67 (0.20)	1
Loadings	PW1	0.52 (0.05)	0.62	0.52 (0.05)	0.71
	PW2	0.47 (0.03)	0.53	0.47 (0.03)	0.63
	PW3	0.49 (0.04)	0.52	0.49 (0.04)	0.61
	PW4	0.54 (0.04)	0.44	0.54 (0.04)	0.54
	PW5	0.60 (0.06)	0.59	0.60 (0.06)	0.68
	NW1	-0.62 (0.10)	-0.50	0.62 (0.10)	0.60
	NW2	-0.57 (0.07)	-0.46	0.57 (0.07)	0.56
	NW3	-0.63 (0.20)	-0.70	0.63 (0.20)	0.79
	NW4	-0.46 (0.03)	-0.35	0.46 (0.03)	0.43
	NW5	-0.64 (0.16)	-0.62	0.64 (0.16)	0.71
Intercepts	PW1	4.34 (0.05)		4.34 (0.05)	
	PW2	4.13 (0.05)		4.13 (0.05)	
	PW3	4.06 (0.05)		4.06 (0.05)	
	PW4	3.95 (0.06)		3.95 (0.06)	
	PW5	4.10 (0.06)		4.10 (0.06)	
	NW1	2.30 (0.07)		3.92 (0.44)	
	NW2	2.03 (0.07)		3.66 (0.35)	
	NW3	1.55 (0.11)		4.21 (0.39)	
	NW4	2.62 (0.04)		3.85 (0.25)	
	NW5	1.62 (0.10)		3.44 (0.55)	
Residual Variances	PW1	0.43 (0.02)		0.43 (0.02)	
	PW2	0.56 (0.02)		0.56 (0.02)	
	PW3	0.65 (0.02)		0.65 (0.02)	
	PW4	1.20 (0.05)		1.20 (0.05)	
	PW5	0.70 (0.04)		0.70 (0.04)	
	NW1	1.15 (0.04)		1.15 (0.04)	
	NW2	1.20 (0.05)		1.20 (0.05)	
	NW3	0.40 (0.03)		0.40 (0.03)	
	NW4	1.56 (0.05)		1.56 (0.05)	
	NW5	0.67 (0.05)		0.67 (0.05)	

Note. The standard deviation of unstandardized model results is reported in parentheses. The results presented in this table were previously reported in Steinmann, Strietholt, et al. (2022)'s Table 4 which used the same sample. Sample size n = 4, 938.

Figure 3

Model-implied Item Intercorrelations Across Positively and Negatively Worded Items



Note. Negatively worded items were not reverse-coded. Weighted sample size for the consistent class n_1 = 4,370; Weighted sample size for the inconsistent class n_2 = 568.

predictors matter not only on their own but also when considered simultaneously, with the highest standardized regression coefficient for reading comprehension ability. The model with both abilities and personalities had the lowest AIC and BIC and was the preferred model according to both AIC and BIC weights (both were 100%). The log-likelihood ratio test which compares the other nested models with this preferred model also supported that this model fitted significantly better than the null model or models with only abilities or personalities ($p < 0.01$ for all three comparisons).

The negative coefficients of abilities implied that students with higher abilities were less likely to be (classified as) an inconsistent responder on the self-esteem scale. Cognitive speed was the only ability predictor not showing a significant relation latent class membership in all the models, which is not in line with the expectation. One conjecture is that the measure

Table 7

Three-step Logistic Models Predicting Membership to the Latent Class of Inconsistent Responders

		Models with Single predictor b (SE)	Ability Model b (SE)	Personality Model b (SE)	Full Model b (SE)
	Intercept		**-2.08 (0.09)**	**-1.97 (0.09)**	**-2.19 (0.10)**
Ability	Reading Comprehension	**-0.79 (0.10)**	**-0.68 (0.10)**		**-0.67 (0.10)**
	Reading Speed	**-0.45 (0.11)**	-0.13 (0.10)		-0.11 (0.10)
	Cognitive Reasoning	**-0.41 (0.08)**	-0.14 (0.08)		**-0.17 (0.08)**
	Cognitive Speed	-0.13 (0.09)	-0.01 (0.08)		-0.01 (0.09)
Personality	Conscientiousness	**-0.43 (0.07)**		**-0.41 (0.08)**	**-0.42 (0.08)**
	Extraversion	**-0.30 (0.07)**		**-0.24 (0.07)**	**-0.23 (0.08)**
	Neuroticism	**0.31 (0.07)**		**0.25 (0.07)**	**0.22 (0.07)**
	Agreeableness	**-0.16 (0.08)**		-0.04 (0.09)	-0.10 (0.09)
	Openness	-0.06 (0.08)		0.00 (0.09)	0.08 (0.09)

Note. Coefficients in bold are statistically different from zero at the 5% significance level. Ability model: model with four ability predictors; Personality model: model with five personality predictors; Full model: model with all four ability and all five personality predictors. The personality and ability predictors were z-standardized. Sample size $n = 4,938$.

Table 8

Comparing Logistic Models Predicting Membership to the Latent Class of Inconsistent Responders

	Null Model	Ability Model	Personality Model	Full Model
-Log-likelihood	1761 (0)	1675 (2)	1705 (4)	**1623 (4)**
AIC	3525 (0)	3361 (3)	3423 (8)	**3266 (9)**
BIC	3531 (0)	3393 (3)	3462 (8)	**3331 (9)**

Note. Null model: model without predictors; Ability model: model with four ability predictors; Personality model: model with five personality predictors; Full model: model with both four ability and five personality predictors. In parentheses, the standard deviation across the analyses of the multiple imputed datasets is reported for each of the fit measures. Sample size $n = 4,938$.

of cognitive speed might be content-wise weak because there is uncertainty regarding what is being measured in the Digit-Symbol Test (see e.g., Jaeger, 2018).

In terms of personalities, students who self-reported being more conscientious, extraverted and agreeable were less likely to be an inconsistent responder, while students who self-reported being more neurotic were more likely to be an inconsistent responder. Openness was the only personality predictor not showing a significant relation with latent class membership in all the models.

Taking all covariates into account, reading comprehension was the dominant predictor of membership to the latent class of inconsistent responders. Students, having scored 2 standard deviations below the average reading comprehension score ($Z = -2$), were estimated to have a probability of 30% (i.e., $Pr(Y = 1|X = -2) = exp(-2.19 + (-2)(-0.67))$ $/(1 + exp(-2.19 + (-2)(-0.67)))$) of having been classified as an inconsistent responder, while this probability was 10% or 3% for students with an average ($Z = 0$) or a high reading score ($Z = 2$), given all the other covariates at the average level. Conscientiousness was also a significant predictor, though to a lesser extent. For conscientiousness, the corresponding estimated probabilities were 21% (low conscientiousness, $Z = -2$), 10% (average conscientiousness, $Z = 0$), and 5% (high conscientiousness, $Z = 2$). Extraversion, neuroticism, and cognitive reasoning contributed some information as well, although their effects were relatively small. Other predictors did not have any significant unique contribution beyond these factors.

Discussion

This study investigated the roles of cognitive abilities and personality traits in the inconsistent responding phenomenon on a mixed-worded self-esteem scale. In doing so, the present study has three primary contributions. First, it is the first study in the surveyed literature which tests the two competing sets of individual determinants of inconsistent responders simultaneously, and thus the study addresses this research gap. Second, the study extends the limited existing empirical research on the relationship between personality traits and inconsistent responding from an individual-centered perspective. Third, the findings

enhance the understanding of the characteristics that are more prevalent in individuals who exhibit inconsistent responses and have implications for the use of mixed-worded scales.

The respondents were classified into the consistent and the inconsistent group using a factor mixture model. The model identified 11% of students as inconsistent responders, indicating that they tended to respond similarly to both PW and NW items on the scale. The estimated class memberships of students were further related to their cognitive abilities and personality traits with logistic regression models including different sets of covariates. Consistent with previous studies, students with lower reading and cognitive abilities were more likely to be classified as inconsistent responders (cf. Bolt et al., 2020; Marsh, 1986; Steedle et al., 2019; Steinmann, Strietholt, et al., 2022). As expected, a lower self-reported conscientiousness level was associated with higher probabilities of being classified as an inconsistent responder. In addition, students who self-reported to be less extraverted and more neurotic were more likely to be classified as inconsistent responders, although no prior expectations of their directions were made considering the limited state of previous theoretical work and empirical research. However, it is not expected that cognitive speed was not associated with inconsistent class membership in all models. This might be because the Digit-Symbol Test used in NEPS is non-verbal and a relatively weak measure regarding its content.

The initial motivation of the study was to compare two theories for the inconsistent response behavior, which were a lack of cognitive abilities versus more of certain personality traits (e.g., inattentiveness and neuroticism). The findings support the important role of cognitive ability, particularly reading comprehension competency as well as the potential effects of personality traits. The model comparison results suggested that both ability and personality blocks have significant contributors to inconsistent responding, with reading comprehension as the strongest predictor followed by conscientiousness. This is in line with the prior expectations and Baumgartner et al.'s (2018) findings that inconsistent responding depends on not only the attention level but also the ability to process the mixed-worded items correctly, and where lack of ability may be a more important cause for response inconsistency than lack of attention.

While Steinmann, Strietholt, et al. (2022) observed a negative correlation between reading ability and inconsistent responding in NEPS Grade 9 students, they did not find such a correlation between conscientiousness and inconsistent responding. In contrast, this study identified a negative correlation between conscientiousness and inconsistent responding in NEPS Grade 5 students. One conjecture for the different findings could be due to the fact that conscientiousness is measured by only two items and through self-reports, which may also be influenced by social desirability bias. Nevertheless, age could be a potential confounding factor and should be considered in future research.

Implication and Generalizability

The findings have implications for survey design and administration. Although incorporating mixed-worded items can reduce the tendency towards acquiescent response styles and provide the opportunity to conduct a check of acquiescence level afterward (Buchholz, 2022), it may pose unintended difficulties and thus lead to inconsistent responses. The difficulty in processing mixed-worded items may be even more of a challenge for young learners who are still developing their reading and cognitive abilities. Therefore, special consideration should be given to wording complexity when designing mixed-worded scales to avoid inconsistent responses due to overly difficult items. Future research may further address if using different types of mixed-worded items (i.e., negated, polar opposite, and reversed items) would influence the difficulties of processing mixed-worded scales and the inconsistent response behavior. Although mixed-worded scales are commonly employed to prompt more attentive responses, their use in a low-stakes context (i.e., the responses do not lead to major consequences for individuals) ironically risks inconsistent responses due to inattentiveness. For the sake of improving the validity and reliability, researchers could consider selecting only PW items when processing response data to mixed-worded scales from young populations or under low-stakes contexts for a robustness check.

The data used in this study is from a representative sample of Grade 5 students in Germany and the study focuses on a particular survey scale on self-esteem. Hence, generalizing the interpretation of the findings to other contexts should be assessed. This

includes but is not limited to other scales (e.g., with less balanced numbers of PW and NW items, or other constructs) or high-stakes situations (in which the respondents are generally more motivated and attentive). One may also speculate that cognitive abilities play a more crucial role in responding to a mixed-worded scale with more complex wording compared to a self-esteem scale, as the former may rely more heavily on cognitive ability than on adequate attentiveness or other factors. In addition, Weems et al. (2003) found that the characteristics of those who responded to PW and NW items most differently had different characteristics across two different scales, implying that the characteristics of the scales also have an impact on the response patterns.

It should also be noted that Germany had a relatively lower prevalence of inconsistent responders compared with other countries in a study analyzing international large-scale assessment data from 37 education systems (Steinmann, Sánchez, et al., 2022). Cross-country differences imply that there might be other extensive factors such as the cultural settings and language issues also affecting the inconsistent responding phenomenon. The language factor is not included in this study due to a lack of variation since students who do not speak German at all (e.g., students who just moved to Germany) are not included in the main survey. However, it could be an interesting direction for future research to explore the impact of language differences (e.g., respondents speaking a language that is more used to having double negations may be less likely to respond inconsistently). Additionally, cultural norms and values may influence the way to interpret mixed-worded items. For example, in some cultures where modesty is considered a virtue, people may be less prone to agree on PW items, while in other cultures, people may be more used to expressing positive attitudes and more likely to agree on PW items.

A limitation of the study is related to the measure of personality traits. The personality measures are relatively poor compared with other measures in the study such as the reading comprehension test. Specifically, self-reported data from students using an 11-item short version of the Big Five personality scale was reported two years after the other variables, and each personality trait was measured based on only two or three items. Moreover, the fact that the Big-Five scale itself is a mixed-worded scale may have led to correlation artifacts. To test

if the findings were robust, a sensitivity check was conducted by selecting only PW or NW items of the personality measures and rerunning the analysis. Even so, the directions of the correlations between personality traits and class memberships remained unchanged, and the key findings remained valid. The results of the sensitivity analysis are presented in Appendix C. Despite our best efforts, it is important to acknowledge that data limitations may have reduced the precision and reliability of the personality measures used in the study. Future studies could consider utilizing more comprehensive and rigorous methods to assess personality, for instance, including parent assessment of personality or using a longer version of the Big Five scale.

In addition to the factor mixture model used in this study, there are other potential methods to identify inconsistent responders in the literature, such as the mean absolute difference (MAD) method (see e.g., Hong et al., 2020; Steedle et al., 2019; Steinmann, Sánchez, et al., 2022). The core of MAD is to calculate the average score difference between the PW items and the reverse-coded NW items, and a larger absolute difference implies a more inconsistent response pattern. A certain threshold needs to be set, and the responders with larger absolute differences exceeding the threshold are marked as inconsistent responders. However, MAD cannot be readily applied to those who have missing responses on mixed-worded items, and establishing the threshold could be challenging under some contexts. As in our case, the midpoint of the response scale is "partly (apply)", which may be perceived as not entirely neutral and instead more towards the "apply" end. Thus, a factor mixture analysis was considered a more appropriate approach in this study. However, future research may consider using multiple ways to identify inconsistent responders and increase the convergent validity of the classification.

Conclusion

In summary, this study examined the roles of cognitive abilities and personality traits in inconsistent response behavior on a mixed-worded self-esteem scale. The findings support that both ability and personality are relevant in identifying inconsistent responders among Grade 5 students in Germany. The strongest contributor to being classified as inconsistent

responders is having a low level of reading comprehension, followed by a low level of conscientiousness. Using mixed-worded scales among young learners with lower reading abilities under low-stakes contexts requires more caution. Future research could consider exploring the impacts of three different types of mixed-worded items and adopting more rigorous measures of personality traits.

References

Asparouhov, T., & Muthén, B. (2014). Auxiliary Variables in Mixture Modeling: Three-Step Approaches Using Mplus. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(3), 329–341. https://doi.org/10.1080/10705511.2014.915181

Asparouhov, T., & Muthén, B. (2022). *Multiple Imputation with Mplus, Version 4*. Mplus Technical Report. http:$%5Cbackslash$$%5Cbackslash$www.statmodel.com

Barnette, J. J. (2000). Effects of Stem and Likert Response Option Reversals on Survey Internal Consistency: If YouFeel 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

Baumgartner, H., Weijters, B., & Pieters, R. (2018). Misresponse to Survey Questions: A Conceptual Framework and Empirical Test of the Effects of Reversals, Negations, and Polar Opposite Core Concepts. *Journal of Marketing Research*, *55*(6), 869–883. https://doi.org/10.1177/0022243718811848

Blossfeld, H.-P., & Roßbach, H.-G. (Eds.). (2019). *Education as a Lifelong Process: The German National Educational Panel Study (NEPS)* (Vol. 3). Springer Fachmedien.

Bolt, D., Wang, Y. C., Meyer, R. H., & Pier, L. (2020). An IRT Mixture Model for Rating Scale Confusion Associated with Negatively Worded Items in Measures of Social-Emotional Learning. *Applied Measurement in Education*, *33*(4), 331–348. https://doi.org/10.1080/08957347.2020.1789140

Borghuis, J., Denissen, J. J. A., Oberski, D. L., Sijtsma, K., Meeus, W. H. J., Branje, S., Koot, H. M., & Bleidorn, W. (2017). Big Five personality stability, change, and co-development across adolescence and early adulthood. *Journal of Personality and Social Psychology*, *113*(4), 641–657. https://doi.org/10.1037/pspp0000138

Buchholz, J. (2022). *Mixed-worded scales and acquiescence in educational large-scale assessments janine buchholz*. OECD Education Working Papers No. 269. https://doi.org/10.1787/8dd310c0-en

DiStefano, C., & Motl, R. W. (2009). Personality correlates of method effects due to negatively worded items on the Rosenberg Self-Esteem scale. *Personality and Individual Differences*, *46*(3), 309–313. https://doi.org/10.1016/j.paid.2008.10.020

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://psycnet.apa.org/fulltext/2000-03696-002.html

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

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.* Bamberg: University of Bamberg, National Education Panel Study.

Gnambs, T., & Schroeders, U. (2020). Cognitive Abilities Explain Wording Effects in the Rosenberg Self-Esteem Scale. *Assessment*, *27*, 404–418. https://doi.org/10.1177/1073191117746503

Gottfredson, L. S. (1997). Mainstream Science on intelligence: An Editorial with 52 Signatories, History and Bibliography. *Intelligence*, *24*, 13–23. https://doi.org/10.1016/S0160-2896(97)90011-8

Haberkorn, K., & Pohl, S. (2013). *Cognitive Basic Skills – Data in the Scientifc Use File*. Bamberg: University of Bamberg, National Education Panel Study.

Hong, M., Steedle, J. T., & Cheng, Y. (2020). Methods of Detecting Insufficient Effort Responding: Comparisons and Practical Recommendations. *Educational and Psychological Measurement*, *80*(2), 312–345. https://doi.org/10.1177/0013164419865316

Jaeger, J. (2018). Digit Symbol Substitution Test: The Case for Sensitivity Over Specificity in Neuropsychological Testing. *Journal of Clinical Psychopharmacology*, *38*(5), 513–519. https://doi.org/10.1097/JCP.0000000000000941

John, O. P., & Srivastava, S. (1999). The Big Five Trait Taxonomy: History, Measurement, and Theoretical Perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of Personality: Theory and Research* (pp. 102–138). Guilford Press.

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

Lang, F. R., Weiss, D., Stocker, A., & von Rosenbladt, B. (2007). Assessing cognitive capacities in computer-assisted survey research: Two ultra-short tests of intellectual ability in the Germany Socio-Economic Panel (SOEP). *Schmollers Jahrbuch. Journal of Applied Social Science Studies*, *127*, 183–192. https://doi.org/10.3790/schm.127.1.183

Marsh, H. W. (1986). Negative item bias in ratings scales for preadolescent children: A cognitive-developmental phenomenon. *Developmental Psychology*, *22*, 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 artifactors? *Journal of Personality and Social Psychology*, *70*, 810–819. https://doi.org/10.1037/0022-3514.70.4.810

Masyn, K. E. (2013). Latent Class Analysis and Finite Mixture Modeling. In T. D. Little (Ed.), *The Oxford Handbook of Quantitative Methods in Psychology: Vol. 2: Statistical Analysis* (pp. 551–611). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199934898.013.0025

Menold, N. (2020). How Do Reverse-keyed Items in Inventories Affect Measurement Quality and Information Processing? *Field Methods*, *32*(2), 140–158. https://doi.org/10.1177/1525822X19890827

Michaelides, M. P. (2019). Negative Keying Effects in the Factor Structure of TIMSS 2011 Motivation Scales and Associations with Reading Achievement. *Applied Measurement in Education*, *32*(4), 365–378. https://doi.org/10.1080/08957347.2019.1660349

Michaelides, M. P., Koutsogiorgi, C., & Panayiotou, G. (2016). Method Effects on an Adaptation of the Rosenberg Self-Esteem Scale in Greek and the Role of Personality Traits. *Journal of Personality Assessment*, *98*(2), 178–188. https://doi.org/10.1080/00223891.2015.1089248

Michaelides, M. P., Zenger, M., Koutsogiorgi, C., Brähler, E., Stöbel-Richter, Y., & Berth, H. (2016). Personality correlates and gender invariance of wording effects in the German version of the Rosenberg Self-Esteem Scale. *Personality and Individual Differences*, *97*, 13–18. https://doi.org/10.1016/j.paid.2016.03.011

Muthén, L. K., & Muthén, B. O. (1998-2017). Mplus User's Guide. Eight Edition.

Paulhus, D. L. (1991). Measurement and Control of Response Bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of Personality and Social Psychological Attitudes* (pp. 17–59). Academic Press. https://doi.org/10.1016/B978-0-12-590241-0.50006-X

Podsakoff, P., MacKenzie, S., Lee, J.-Y., & Podsakoff, N. (2003). Common Method Biases in Behavioral Research: A Critical Review of the Literature and Recommended Remedies. *The Journal of applied psychology*, *88*, 879–903. https://doi.org/10.1037/0021-9010.88.5.879

Pohl, S., & Carstensen, C. H. (2012). *NEPS Technical Report - Scaling the Data of the Competence. Tests (NEPS Working Paper No. 14)*. Bamberg: Otto-Friedrich-Universität, National Education Panel Study.

Quilty, L., Oakman, J., & Risko, E. (2006). Correlates of the Rosenberg Self-Esteem Scale Method Effects. *13*, 99–117. https://doi.org/10.1207/s15328007sem1301_5

R Core Team. (2020). R: A language and environment for statistical computing. https://www.R-project.org/

Raven, J. C. (1941). Standardization of progressive matrices, 1938. *British Journal of Medical Psychology*, *19*, 137–150. https://doi.org/10.1111/j.2044-8341.1941.tb00316.x

Rosenberg, M. (1965). *Society and the Adolescent Self-Image*. Princeton University Press.

Rubin, D. B. (1978). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons.

Schmitt, N., & Stuits, 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

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

Steinhauer, H. W., & Zinn, S. (2016). *NEPS Technical Report for Weighting: Weighting the Sample of Starting Cohort 3 of the National Educational Panel Study (Waves 1 to 5)*. Bamberg: Leibniz Institute for Educational Trajectories, National Education Panel Study. https://doi.org/10.5157/NEPS:SC3:5.0.0

Steinmann, I., Sánchez, D., van Laar, S., & Braeken, J. (2022). The impact of inconsistent responders to mixed-worded scales on inferences in international large-scale assessments. *Assessment in Education: Principles, Policy & Practice*, *29*(1), 5–26. https://doi.org/10.1080/0969594X.2021.2005302

Steinmann, I., Strietholt, R., & Braeken, J. (2022). A constrained factor mixture analysis model for consistent and inconsistent respondents to mixed-worded scales. *Psychological Methods*, *27*(4), 667–702. https://doi.org/10.1037/met0000392

Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, *11*(1), 192–196. https://doi.org/10.3758/BF03206482

Wang, W.-C., Chen, H.-F., & Jin, K.-Y. (2015). Item response theory models for wording effects in mixed-format scales. *Educational and Psychological Measurement*, *75*(1), 157–178. https://doi.org/10.1177/0013164414528209

Weems, H. G., Onwuegbuzie, A., Schreiber, J., & Eggers, S. (2003). Characteristics of respondents who respond differently to positively and negatively worded items on rating scales. *Assessment & Evaluation in Higher Education*, *28*, 587–606. https://doi.org/10.1080/0260293032000130234

Weijters, B., & Baumgartner, H. (2012). Misresponse to Reversed and Negated Items in Surveys: A Review. *Journal of Marketing Research*, *49*(5), 737–747. https://doi.org/10.1509/jmr.11.0368

Zimmermann, S., Gehrer, K., Artelt, C., & Weinert, S. (2012). *The Assessment of Reading Speed in Grade 5 and Grade 9 Status. Scientifc Use File 2012*. Bamberg: University of Bamberg, National Education Panel Study.

Appendix A

GDPR Documentation & Ethical Approval

The current study used anonymous data and followed proper protocol with regard to GDPR (Data Protection Regulation). The NSD (Norsk Senter for Forskningsdata) GDPR notification test and NSD GDPR test outcome are presented below.

Denne siden på norsk (/personvernombud/meld_prosjekt/meldeplikttest.html)

Will you be processing personal data?

Are you unsure whether your project is subject to notification? Feel free to try our informal Notification test. Note that the test is intended as a guidance and is not a formal assessment.

Will you be collecting/processing directly identifiable personal data? ○ Yes ◉ No

A person will be directly identifiable through name, social security number, or other uniquely personal characteristics.

Read more about personal data (/personvernombud/en/help/vocabulary.html?id=8) and notification (/personvernombud/en/notify/index.html).

NB! Even though information is to be anonymized in the final thesis/report, check the box if identifying personal data is to be collected/processed in connection with the project.

Will directly identifiable personal information be linked to the data (e.g. through a reference number which refers to a separate list of names/scrambling key)? ○ Yes ◉ No

Note that the project will be subject to notification even if you cannot access the scrambling key (/personvernombud/en/help/vocabulary.html?id=11), as the procedure often is when using a data processor (/personvernombud/en/help/vocabulary.html?id=3), or in register-based studies (/personvernombud/en/help/research_methods/register_studies.html).

Will you be collecting/processing background information that may identify individuals (indirectly identifiable personal data)? ○ Yes ◉ No

A person will be indirectly identifiable if it is possible to identify him/her through a combination of background information (such as place of residence or workplace/school, combined with information such as age, gender, occupation, etc.).

Will there be registered personal data (directly/indirectly/via IP or email address, etc.) using online surveys? ○ Yes ◉ No

Please note that the project will be subject to notification even if you as a student/researcher cannot access the link to the IP or email address, as the procedure often is when using a data processor.

Read more about online surveys (/personvernombud/en/help/research_methods/online_surveys.html).

Will there be registered personal data using digital photo or video files? ○ Yes ◉ No

Photo/video recordings of faces will be regarded as identifiable personal data. In order for a voice to be considered as identifiable, it must be registered in combination with other background information, in such a way that people can be recognized.

Show results

Notify project

Do I have to notify my project? (/personvernombud/en/notify/index.html)

Notification Form (/personvernombud/en/notify/meldeskjema_link)

Notifying changes (/personvernombud/en/notify/notifying_changes.html)

Get help notifying your project

Processing the notification (/personvernombud/en/help/index.html)

Frequently asked questions (/personvernombud/en/help/faq.html)

Vocabulary (/personvernombud/en/help/vocabulary.html)

Research topics (/personvernombud/en/help/research_topics/)

Research methods (/personvernombud/en/help/research_methods/)

Information and consent (/personvernombud/en/help/information_consent/)

Other approvals (/personvernombud/en/help/other_approvals/)

Result of Notification Test: Not Subject to Notification

You have indicated that neither directly or indirectly identifiable personal data will be registered in the project.

If no personal data is to be registered, the project will not be subject to notification, and you will not have to submit a notification form.

Please note that this is a guidance based on information that you have given in the notification test and not a formal confirmation.

For your information: *In order for a project not to be subject to notification, we presuppose that all information processed using electronic equipment in the project remains anonymous.*

Anonymous information is defined as information that cannot identify individuals in the data set in any of the following ways:

- *directly, through uniquely identifiable characteristic (such as name, social security number, email address, etc.)*
- *indirectly, through a combination of background variables (such as residence/institution, gender, age, etc.)*
- *through a list of names referring to an encryption formula or code, or*
- *through recognizable faces on photographs or video recordings.*

Furthermore, we presuppose that names/consent forms are not linked to sensitive personal data.

Kind regards,
NSD Data Protection

Appendix B

Data Management and Analysis Code

NEPS data is not publicly available. The basic requirement for any NEPS data access is the conclusion of a Data Use Agreement with the Leibniz Institute for Educational Trajectories. For more information, see https://www.neps-data.de/Data-Center/Data-Access.

R and Mplus syntaxes related to this master thesis can be found via the link: https://drive.google.com/drive/folders/1Uc7OYh18GQbbCVYuXQOf6oyu1Nn7rRD5?usp=share_link. Specifically, the following parts can be found:

- Data management and descriptives: DATA_DESC.R
- Multiple imputation: MI.inp
- Factor model analysis (also used to generate imputed datasets with latent class memberships for further one-step logistic regression): FMA.inp
- Three-step logistic regression (an example of the full model including both ability and personality predictors): MODEL_3STEP.inp
- One-step logistic regression (an example of the full model including both ability and personality predictors): MODEL_1STEP.inp
- Data preparation for sensitivity analysis using only positively- or negatively worded items of the personality measures: SEN_CHECK.R

Appendix C

Supplemental Material

Table C1

Three-step Logistic Models Predicting Membership to the Latent Class of Inconsistent Responders (Personality Predictors Measured by Only Positively Worded Items)

		Models with Single predictor b (SE)	Ability Model b (SE)	Personality Model b (SE)	Full Model b (SE)
	Intercept		**-2.08 (0.09)**	**-1.96 (0.09)**	**-2.18 (0.09)**
Ability	Reading Comprehension	**-0.79 (0.10)**	**-0.68 (0.10)**		**-0.65 (0.10)**
	Reading Speed	**-0.45 (0.11)**	-0.13 (0.10)		-0.10 (0.10)
	Cognitive Reasoning	**-0.41 (0.08)**	-0.14 (0.08)		-0.15 (0.08)
	Cognitive Speed	-0.13 (0.09)	-0.01 (0.08)		-0.01 (0.09)
Personality	Conscientiousness	**-0.37 (0.07)**		**-0.33 (0.07)**	**-0.32 (0.07)**
	Extraversion	**-0.31 (0.07)**		**-0.23 (0.08)**	**-0.19 (0.08)**
	Neuroticism	**0.32 (0.08)**		**0.31 (0.08)**	**0.28 (0.08)**
	Agreeableness	**-0.17 (0.08)**		-0.06 (0.08)	-0.07 (0.08)
	Openness	-0.11 (0.08)		-0.05 (0.09)	0.03 (0.08)

Note. Coefficients in bold are statistically different from zero at the 5% significance level. Ability model: model with four ability predictors; Personality model: model with five personality predictors; Full model: model with all four ability and all five personality predictors. The personality and ability predictors were z-standardized. Sample size $n = 4,938$.

Table C2

Comparing Logistic Models Predicting Membership to the Latent Class of Inconsistent Responders (Personality Predictors Measured by Only Positively Worded Items)

	Null Model	Ability Model	Personality Model	Full Model
-Log-likelihood	1761 (0)	1676 (1)	1708 (5)	**1635 (4)**
AIC	3525 (0)	3361 (3)	3428 (10)	**3290 (8)**
BIC	3531 (0)	3394 (3)	3467 (10)	**3355 (8)**

Note. Null model: model without predictors; Ability model: model with four ability predictors; Personality model: model with five personality predictors; Full model: model with both four ability and five personality predictors. In parentheses, the standard deviation across the analyses of the multiple imputed datasets is reported for each of the fit measures. Sample size $n = 4,938$.

Table C3

Three-step Logistic Models Predicting Membership to the Latent Class of Inconsistent Responders (Personality Predictors Measured by Only Negatively Worded Items)

		Models with Single predictor b (SE)	Ability Model b (SE)	Personality Model b (SE)	Full Model b (SE)
	Intercept		**-2.08 (0.09)**	**-1.90 (0.09)**	**-2.13 (0.09)**
Ability	Reading Comprehension	**-0.80 (0.10)**	**-0.69 (0.10)**		**-0.70 (0.10)**
	Reading Speed	**-0.45 (0.11)**	-0.13 (0.10)		-0.12 (0.10)
	Cognitive Reasoning	**-0.41 (0.08)**	-0.13 (0.08)		-0.15 (0.08)
	Cognitive Speed	-0.13 (0.09)	-0.01 (0.08)		-0.01 (0.08)
Personality	Conscientiousness	**-0.31 (0.08)**		**-0.32 (0.09)**	**-0.33 (0.09)**
	Extraversion	**-0.15 (0.08)**		-0.14 (0.08)	**-0.16 (0.08)**
	Neuroticism	0.15 (0.08)		**0.16 (0.08)**	**0.14 (0.07)**
	Agreeableness	-0.09 (0.07)		-0.01 (0.07)	-0.10 (0.08)
	Openness	-0.03 (0.07)		0.07 (0.07)	0.11 (0.08)

Note. Coefficients in bold are statistically different from zero at the 5% significance level. Ability model: model with four ability predictors; Personality model: model with five personality predictors; Full model: model with all four ability and all five personality predictors. The personality and ability predictors were z-standardized. Sample size $n = 4,938$.

Table C4

Comparing Logistic Models Predicting Membership to the Latent Class of Inconsistent (Personality Predictors Measured by Only Negatively Worded Items)

	Null Model	Ability Model	Personality Model	Full Model
-Log-likelihood	1761 (0)	1675 (1)	1737 (3)	**1646 (5)**
AIC	3525 (0)	3360 (2)	3485 (7)	**3311 (9)**
BIC	3531 (0)	3392 (2)	3524 (7)	**3376 (9)**

Note. Null model: model without predictors; Ability model: model with four ability predictors; Personality model: model with five personality predictors; Full model: model with both four ability and five personality predictors. In parentheses, the standard deviation across the analyses of the multiple imputed datasets is reported for each of the fit measures. Sample size $n = 4,938$.

Table C5

Original German Item Wording of the Self-esteem Scale in NEPS Starting Cohort 3, Wave 1 (Grade 5)

Item	Inwieweit treffen folgende Aussagen auf dich zu?
PW1	Alles in allem bin ich mit mir selbst zufrieden.
NW1	Hin und wieder denke ich, dass ich gar nichts tauge.
PW2	Ich besitze eine Reihe guter Eigenschaften.
PW3	Ich kann vieles genauso gut wie die meisten anderen Menschen auch.
NW2	Ich fürchte, es gibt nicht viel, worauf ich stolz sein kann.
NW3	Ich fühle mich von Zeit zu Zeit richtig nutzlos.
PW4	Ich halte mich für einen wertvollen Menschen, jedenfalls bin ich nicht weniger wertvoll als andere auch.
NW4	Ich wünschte, ich könnte vor mir selbst mehr Achtung haben.
NW5	Alles in allem neige ich dazu, mich für eine Versagerin oder einen Versager zu halten.
PW5	Ich habe eine positive Einstellung zu mir selbst gefunden.

Note. PW represents positively worded items; NW represents negatively worded items. Response scale: trifft gar nicht zu = 1; trifft eher nicht zu = 2; teils/teils = 3; trifft eher zu = 4; trifft völlig zu = 5. The items are ordered in the original sequence as presented in the questionnaire.

Table C6

Original German Item Wording of the Big-Five Scale in NEPS Starting Cohort 3, Wave 3 (Grade 7)

Inwieweit treffen die folgenden Aussagen auf dich zu?	Trait
a) Ich bin eher zurückhaltend, reserviert.	Extraversion
b) Ich schenke anderen leicht Vertrauen, glaube an das Gute im Menschen.	Agreeableness
c) Ich bin bequem, neige zur Faulheit.	Conscientiousness
d) Ich bin entspannt, lasse mich durch Stress nicht aus der Ruhe bringen.	Neuroticism
e) Ich habe nur wenig künstlerisches Interesse.	Openness
f) Ich gehe aus mir heraus, bin gesellig.	Extraversion
g) Ich neige dazu, andere zu kritisieren.	Agreeableness
h) Ich erledige Aufgaben gründlich.	Conscientiousness
i) Ich werde leicht nervös und unsicher.	Neuroticism
j) Ich habe eine aktive Vorstellungskraft, bin phantasievoll.	Openness
k) Ich bin rücksichtsvoll zu anderen, einfühlsam.	Agreeableness

Note. Response scale: trifft gar nicht zu = 1; trifft eher nicht zu = 2; teils/teils = 3; trifft eher zu = 4; trifft völlig zu = 5. The items are ordered in the original sequence as presented in the questionnaire.