Personality and Individual Differences 222 (2024) 112573

Contents lists available at ScienceDirect

Personality and Individual Differences

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

Competing explanations for inconsistent responding to a mixed-worded self-esteem scale: Cognitive abilities or personality?

Jianan Chen [a,*], Isa Steinmann [b], Johan Braeken [a]

[a] CEMO: Centre for Educational Measurement, University of Oslo, Oslo, Norway
[b] Department of Primary and Secondary Teacher Education, Oslo Metropolitan University, Oslo, Norway

ARTICLE INFO

Keywords:
Inconsistent responding
Mixed-worded scale
Rosenberg's self-esteem scale
Cognitive abilities
Big Five personality traits
Factor mixture model
NEPS

ABSTRACT

In survey scale design, a mixed-worded format intends to ensure attentiveness as respondents need to take into account the wording direction when answering an item. However, some respondents tend to deliver inconsistent responses (i.e., agreeing or disagreeing with both positively and negatively-worded items), posing a validity concern. Two potential directions driving inconsistent responding have each individually been put forward: lack of cognitive abilities to effectively deal with the mixed-wording and sheer carelessness while responding. Using a factor mixture approach, we investigated inconsistent responding on Rosenberg's self-esteem scale as a function of both cognitive ability and personality. Among n = 4938 Grade 5 students from the German National Educational Panel Study (NEPS), 11 % were classified as inconsistent respondents and 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). Model comparison results indicated that both ability and personality predictors matter with a more prominent role for ability (especially reading comprehension). We discuss the implications of these findings for mixed-wording's suitability for scale construction and different populations and how researchers can deal with suspected inconsistent respondents in their survey data.

1. Introduction

A mixed-worded scale measures a target construct by means of items with both positive and negative wording. The mixed wording can be generated by using negations (e.g., "no", "not", "un-", "non-") or antonyms (e.g., "happy" vs. "sad") (Menold, 2020). For instance, Rosenberg (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". For a mixed-worded scale, respondents are required to attentively read each item and subsequently switch sides of the response scale according to the wording direction of the item. For example, to express a high self-esteem level, a consistent respondent is expected to agree on PW items and disagree on NW items of the self-esteem scale.

Mixed-worded scales have gained popularity in survey design as they enable a logical quality assurance check of the consistency of the respondents (Huang, Curran, Keeney, Poposki, & DeShon, 2012) and can reduce acquiescence response bias (the tendency that respondents are willing to agree rather than disagree on an item regardless of its content) (Paulhus, 1991). However, a body of literature —from an instrument-centered perspective using factor analysis and from an individual-centered perspective focusing on individuals with differential response patterns— showed that mixed-worded scales should be used with caution since they could lead to unintended consequences, threatening the validity of survey data and conclusions of research conducted using these data (e.g., Marsh, 1996; Steedle, Hong, & Cheng, 2019; Steinmann, Sánchez, van Laar, & Braeken, 2022).

From an instrument-centered perspective, the intercorrelations across PW and NW items are often attenuated; PW and NW items are less negatively correlated than expected under a unidimensional scale (Dunbar, Ford, Hunt, & Der, 2000; Marsh, 1986; Steinmann, Strietholt, & Braeken, 2022). The use of mixed-worded scales risks lower reliability (e.g., Barnette, 2000; Steinmann, Sánchez, et al., 2022) and poor model fit of a unidimensional factor model (e.g., Marsh, 1996; Steinmann, Sánchez, et al., 2022). Introducing method factors for PW items and/or NW items would typically improve model fit (e.g., DiStefano & Motl, 2009; Wang, Chen, & Jin, 2015). Hence, mixed-worded scales lead to more complex latent structures than intended. This method effect of the

* Corresponding author.
E-mail address: jianan.chen@cemo.uio.no (J. Chen).

https://doi.org/10.1016/j.paid.2024.112573
Received 15 November 2023; Received in revised form 27 December 2023; Accepted 27 January 2024

wording or keying on the item dependence structure has been traditionally regarded as causing construct irrelevant variance, implying that it should be eliminated or at least minimized (Marsh, 1996).

Furthermore, the method factors correlate with respondent characteristics such as cognitive abilities (e.g., Dunbar et al., 2000; Gnambs & Schroeders, 2020; Marsh, 1986), and conscientiousness and neuroticism (e.g., Michaelides et al., 2016; Michaelides, Koutsogiorgi, & Panayiotou, 2016; Quilty, Oakman, & Risko, 2006). These findings imply that the wording effect might not equally affect all respondents and that an alternative, individual-centered perspective can offer a fruitful complement to the more instrument-centered factor-analytical studies.

Instead of searching for a common wording-method effect, studies from an individual-centered perspective have aimed at identifying inconsistent respondents who do not switch sides of the response scale following the wording direction. An inconsistent respondent would for instance strongly agree with both PW and NW items, with a resulting lack of internal consistency in their responses across items of the mixed-worded scale. About 10–20 % of respondents are typically flagged as delivering inconsistent item responses (e.g., Kam & Chan, 2018; Steedle et al., 2019; Steinmann, Sánchez, et al., 2022; Steinmann, Strietholt, & Braeken, 2022).

Research into interindividual differences in inconsistent responding has focused on either cognitive abilities or personality traits as key factors. Both directions have a common starting point in that mixed-worded scales are considered challenging to some respondents in a particular fashion. Lower cognitive abilities are seen as a risk factor that makes respondents more prone to making inconsistency mistakes due to misreading and/or misinterpreting items, especially concerning wording changes across items. Reading ability, cognitive reasoning, and academic competence (e.g., high school grade point average) were found to be negatively correlated with a higher risk of inconsistent responding (e.g., Bolt, Wang, Meyer, & Pier, 2020; Marsh, 1986; Steedle et al., 2019; Steinmann, Strietholt, & Braeken, 2022).

Personality traits have been put forward as another major source of interindividual differences in inconsistent responding, with conscientiousness and neuroticism correlating significantly to method factors of mixed-worded scales (Michaelides, Zenger, et al., 2016; Quilty et al., 2006; Schmitt & Stuits, 1985). Some logical theoretical conjectures can be made. First, certain personality traits could play a role in detecting wording differences; e.g., conscientiousness measures an individual's inclination to be attentive and comply and was found to be negatively related to insufficient effort responding (Bowling et al., 2016), we may therefore speculate that a more conscientious respondent will go through the questionnaire more carefully, enabling them to notice changes in item wording in a mixed-worded scale and respond consistently. Second, individuals' personality traits could relate to how they would attend to and interpret negatively- versus positively-worded items; e.g. in an electroencephalography study, individuals with higher neuroticism and lower extraversion tended to demonstrate more sustained attention and in-depth processing of negative information, and those with higher neuroticism and higher extraversion tended to identify positive word-content earlier on (Ku, Chan, & Lai, 2020). However, it is not directly obvious how to translate these observed correlations between, on the one side the latter event-related potentials or the former method factor (i.e., residual common item response variation not due to the target construct) and on the other side personality traits, into risk/protective characteristics for inconsistent responding. Given the scarce literature on how these individual differences in personality impact response consistency, further exploration is needed to understand the relationship between personality and inconsistent responding.

1.1. The present study

Studies on inconsistent responding from an individual-centered perspective that include both personality traits and cognitive ability appear to be missing from the literature. Using data from the German National Educational Panel Study (NEPS), the present study investigates whether individual differences in cognitive ability and Big-5 personality traits are associated with inconsistent responding on Rosenberg's (1965) popular mixed-worded self-esteem scale. The study has three research objectives: (i) a replication test of the negative correlation between low cognitive ability and inconsistent responding found in the literature, and this for a range of cognitive ability measures covering cognitive reasoning, cognitive speed, reading comprehension and reading speed; (ii) an exploration, from an individual-centered perspective, of the relation between Big-5 personality traits and inconsistent responding found in the method-factor literature; and (iii) a competitive comparison between cognitive abilities and personality traits concerning their relation to inconsistent responding.

2. Method

The National Educational Panel Study (NEPS) is a large-scale educational study in Germany providing longitudinal data on individual educational processes and outcomes (Blossfeld & Rossbach, 2019). The three core elements for our research objectives —a mixed-worded scale, and measures of cognitive ability and personality traits— are present in NEPS.

2.1. Sample

The NEPS Starting Cohort 3 (SC3) was used, which targeted Grade 5 students in Germany during the 2010/2011 academic year. 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, were excluded from NEPS (Blossfeld & Rossbach, 2019). NEPS followed a two-stage stratified cluster sampling procedure with schools as the primary first-stage sampling units and up to two Grade 5 classes randomly selected per school in the second stage, see the NEPS technical reports (NEPS Network, 2022) for full details.

Our sample included all students who participated in the first wave, except those from special needs schools or those who were part of the oversampling for migrant students (these groups received a different survey). Additionally, 34 students were excluded from the study as they did not provide any responses (i.e., with all ten items missing) on the mixed-worded scale. The effective sample contained $n = 4972 - 34 = 4938$ Grade 5 students (age in years $M = 11.04$, $SD = 0.64$) from 203 schools and approximately 50 % of them were girls.

2.2. 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 ability tests were extracted from Wave 1 (Grade 5 in 2010/2011) and their self-reported Big-5 personality traits were extracted from Wave 3 (Grade 7 in 2012/2013). Full operationalization details of these measures can be found in the NEPS technical reports (NEPS Network, 2022), but we provide a summary below. 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-5 personality trait scores from Wave 3. Among the 4938 students in Wave 1, 684 (14 %) students did not participate in Wave 3 (due to e.g., switching or repeating schools).

2.2.1. Mixed-worded scale: self-esteem

The German version of the Rosenberg Self-esteem Scale (Rosenberg, 1965) forms the basis for our study of inconsistent respondents. Ten five-point Likert items (variables: 't66003a'-'t66003j'), five positively-worded and five negatively-worded, intended to measure an

individual's favorable or unfavorable self-perception (see Table 1).

2.2.2. Outcome variable: classification as inconsistent respondent

To classify a student as a consistent or an inconsistent respondent on the self-esteem scale, the constrained factor mixture model proposed by Steinmann, Strietholt, and Braeken (2022) was adopted (see Fig. 1). The model assumed the existence of two latent classes within the target population, namely a consistent class and an inconsistent class. In the consistent class, positively and negatively-worded items have opposite-sign factor loadings (λ's) reflecting the implied switching of response scale with the direction of the items' wording, whereas this is lacking in the inconsistent class (i.e., same sign across mixed wordings). To keep the same measurement scale across the two classes, the positively-worded items were considered intercept (ν's) and loading invariant across classes, and the negatively-worded items had opposite factor loadings across classes.

Students were classified into the inconsistent and consistent respondent class based on their maximum posterior class membership probability. This binary classification is our core outcome measure. Average class membership probabilities and entropy were used to evaluate classification precision.

2.2.3. Main predictors: 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.

2.2.3.1. Cognitive reasoning. The NEPS reasoning test (NEPS-MAT) is a progressive matrices test measuring non-verbal reasoning. The test consisted of three sets of four items each, with a time limit of three minutes per set. The sumscore correct (with a maximum of 12 points) was recorded as variable 'dgg5_sc3b'.

2.2.3.2. Cognitive speed. The NEPS Picture Symbol Test (NEPS-BZT) measured perceptual speed, reflecting the speed of information processing. The students had to match figures or numbers with graphical symbols as quickly as possible. The test consisted of three sets of 31 items, with a time limit of 30 s per set. The sumscore correct (with a maximum of 93 points) was recorded as variable 'dgg5_sc3a'.

2.2.3.3. Reading comprehension. The reading comprehension test addressed the ability to process written text proficiently in everyday situations, with item formats of multiple-choice, decision-making tasks, and matching tasks. Within 28 min, the students had to complete the test which consisted of five texts (five to seven items per text). A mean-centered model-based scale score was recorded as variable 'reg5_sc1'.

Table 1
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	On the whole, I am satisfied with myself.
NW1	At times I think I am no good at all.
PW2	I feel that I have a number of good qualities.
PW3	I am able to do things as well as most other people.
NW2	I feel I do not have much to be proud of.
NW3	I certainly feel useless at times.
PW4	I feel that I am a person of worth, at least on an equal plane with others.
NW4	I wish I could have more respect for myself.
NW5	All in all, I am inclined to feel that I am a failure.
PW5	I take a positive attitude toward 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. For the original German version, see Appendix C1.

2.2.3.4. Reading speed. The reading speed test aimed to assess the respondents' automatized reading processes; it had 51 items and required the respondents to rate short sentences as either true or false. The reading speed score 'rsg5_sc3' was recorded as the number of correctly judged sentences within the two-minute time limit.

2.2.4. Main predictors: five personality traits

In Wave 3 (i.e., grade 7), NEPS introduced the Big-5 self-reported personality measures in the student survey. Although measured two years after other variables, students' personality traits are considered relatively stable over time (Borghuis et al., 2017). The scale contained 11 items measuring five personality traits (neuroticism, conscientiousness, extraversion, agreeableness, and openness; see Appendix C2). The response scale was a five-point Likert scale, ranging from 1 (Does not apply at all) to 5 (Applies completely). Mean trait scores (variables: 't66800a_g1'-'t66800e_g1') were used as personality measures.

2.3. 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.

In the first step of our analyses, the factor mixture model by Steinmann, Strietholt, and Braeken (2022) was estimated for the self-esteem scale, treating item responses as interval measures, using full information maximum likelihood and an EM algorithm in Mplus (5000 random sets of starting values for the initial estimation stage and 500 optimizations for the final stage).

In the second step of our analyses, we used logistic regression to relate the membership of the inconsistent respondent class to the cognitive ability and personality predictors. Both the uncertainty due to missing information in the predictors and due to estimation error in the latent class membership were accounted for via a multiple imputation approach resulting in a set of 10 imputed datasets. Latent class membership was imputed in line with the individuals' latent class membership probabilities based on the estimated factor mixture model. Predictors were imputed based on a fully saturated model including all predictors and the self-esteem items as auxiliary variables, and treating self-esteem and personality variables as categorical in the imputation model to stay as close to the data as possible. Following a model comparison strategy, 12 models including different sets of covariates (single predictors, ability predictor block, personality predictor block, and a predictor block including ability and personality) were run on these imputed datasets. For each model, the results were combined across imputations following Rubin's (1987) rules. Likelihood ratio tests for nested model comparison, and Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) were reported for the full sets of models.

For all analyses (first and second step), the NEPS 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.

3. Results

3.1. 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 (see Appendix A1). Big-5 Personality trait scores had an average mean of around 3 corresponding to a "partially applies" response.

Between abilities, the correlations were all positive around 0.30, except for correlations of 0.10 between cognitive speed and non-speed

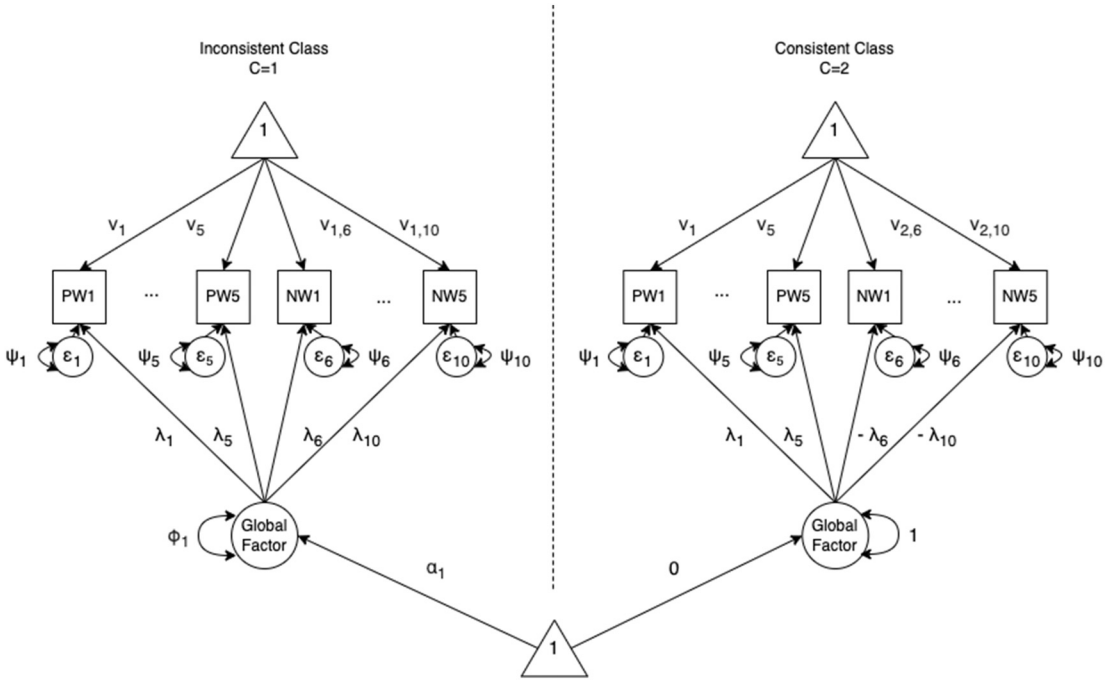

Fig. 1. Constrained factor mixture analysis model to classify in/consistent respondents. *Note.* Representation follows default path diagram conventions. Adapted from Steinmann, Strietholt, and Braeken (2022).

ability measures (see Appendix A2). Between personality facets, the correlations were mostly zero with some exceptions of low positive or negative correlations of around 0.20. Across personalities and abilities, the correlations were close to zero.

3.2. Inconsistent respondents: factor mixture model and classification

The factor mixture model estimated a prevalence of 14 % inconsistent respondents in the population. The congeneric reliability (i.e., coefficient omega) of the self-esteem measure for the consistent class was estimated to be 0.78. More details on the estimated factor mixture model are given in Appendix B1.

3.2.1. Classification

Based on their item response pattern on the self-esteem scale and their corresponding maximum posterior class membership probability, 11 % of the students were assigned to the inconsistent class. The average class membership probability for inconsistent respondents was 0.87 and 0.96 for the consistent respondents, indicating assignments with low uncertainty. The resulting entropy value of 0.83 indicated a crisp classification quality.

To characterize the two resulting classes of the factor mixture model, the observed item means and intercorrelations were computed after classification. For clarity of exposition, we show the wording-aggregated version here in Table 2. Whereas the expected mixed correlation pattern is implied for the consistent respondents class, the inter-item correlations for the inconsistent respondents class were homogeneously positive regardless of the direction of the item wording. For the former inconsistent class, the average item response hardly differed between differently worded items. These results conform to our definition of consistent versus inconsistent respondents on mixed-worded scales.

Table 2
Inter-item correlations and average means across positively- and negatively-worded items.

		Inconsistent class (11 %)		Consistent class (89 %)	
		PW items	NW items	PW items	NW items
Correlation	PW items	0.45		0.34	
	NW items	0.33	0.36	−0.26	0.34
Mean		3.75	3.45	4.10	2.04

Note. Negatively worded items were not reverse-coded. Classification in the in/consistent class is based on maximum posterior class membership and statistics are averaged across items. Sample size $n = 4938$.

3.3. Inconsistent respondent classification as a function of ability and personality

As a single predictor, almost all cognitive ability and personality measures showed a significant relation to inconsistent responding, except for Cognitive Speed and Openness (see Table 3).

3.3.1. Cognitive ability

The negative logistic regression coefficients of the cognitive ability predictors implied that students with higher abilities were less likely to be classified as inconsistent respondents on the self-esteem scale. When considered in block (Ability Model, Table 3), Reading Comprehension was the dominant predictor (Reading Comprehension: OR = 0.66), accounting for most of the relevant predictive variation of the other cognitive ability measures (Reading Speed: OR = 0.92; Cognitive Reasoning: OR = 0.90; Cognitive Speed: OR = 0.99).

3.3.2. Big-5 personality

A lower self-reported conscientiousness level was associated with higher probabilities of being classified as an inconsistent respondent. In addition, students who self-reported to be less extraverted, less agreeable, or more neurotic were more likely to be classified as inconsistent respondents. Openness was the only personality predictor not showing a significant relation with latent class membership. When considered in block (Personality Model, Table 3), Conscientiousness was the dominant predictor (Conscientiousness: OR = 0.73), but the other personality traits retained their predictive sign and relevance, although to a smaller extent (Extraversion: OR = 0.84; Neuroticism: OR = 1.22; Agreeableness: OR = 0.95; Openness: OR = 1.00).

Table 3
Logistic regression models predicting membership to the latent class of inconsistent respondents.

	Single predictor Model	Ability Model	Personality Model	Full Model
	b (SE)	b (SE)	b (SE)	b (SE)
Intercept		**−1.94 (0.06)**	**−1.92 (0.07)**	**−2.01 (0.06)**
Reading comprehension	**−0.49 (0.07)**	**−0.42 (0.07)**		**−0.41 (0.07)**
Reading speed	**−0.27 (0.09)**	−0.08 (0.09)		−0.07 (0.09)
Cognitive reasoning	**−0.29 (0.06)**	−0.10 (0.06)		−0.12 (0.06)
Cognitive speed	−0.09 (0.07)	−0.01 (0.07)		−0.01 (0.07)
Conscientiousness	**−0.34 (0.06)**		**−0.31 (0.07)**	**−0.32 (0.07)**
Extraversion	**−0.22 (0.06)**		**−0.17 (0.07)**	**−0.16 (0.07)**
Neuroticism	**0.25 (0.05)**		**0.20 (0.06)**	**0.18 (0.06)**
Agreeableness	**−0.14 (0.06)**		−0.05 (0.07)	−0.09 (0.07)
Openness	−0.04 (0.06)		0.00 (0.06)	0.05 (0.06)

Note. Coefficients in bold are statistically different from zero at the 5 % significance level. Full model: model with all ability and personality predictors. The personality and ability predictors were standardized. Sample size $n = 4938$.

3.3.3. *Full Model*

The model including both abilities and personalities had the lowest AIC and BIC (see Table 4) and also the log-likelihood ratio tests indicated that this comprehensive model fitted significantly better than the other candidate models ($p < 0.01$ for all comparisons).[1] These model comparison results suggest that it is not an either-or story, but that both cognitive ability and personality uniquely relate to inconsistent responding.

Table 4
Comparing logistic regression models predicting membership to the latent class of inconsistent respondents.

	Null model	Ability model	Personality model	Full model
-Log-likelihood	1970 (26)	1899 (23)	1912 (25)	**1844 (22)**
AIC	3941 (51)	3808 (46)	3835 (51)	**3707 (44)**
BIC	3948 (51)	3825 (46)	3874 (51)	**3772 (44)**

Note. Full model: model with all ability and 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 $n = 4938$.

[1] As a sensitivity check, we used two different methods to classify inconsistent respondents, and the general finding held. First, using a mean absolute difference (MAD) between PW and NW items to quantify the degree of inconsistent responding (e.g., Steedle et al., 2019; Steinmann, Sánchez, et al., 2022) and a threshold of $MAD = 1.5$, 10 % of the students were classified as inconsistent respondents. Second, using a less constrained mixture factor model proposed by Kam and Cheung (2023), 36 % of the students were classified as inconsistent. No matter which method was used, the model with both ability and personality was preferred, with Reading Comprehension and Conscientiousness as the dominant predictors.

An "average student" (i.e., with an average score on all predictors, $X = 0$) would be expected to have a probability of being classified as an inconsistent respondent of 12 % (i.e., inverse logit of the intercept −2.01) in the final model including both predictor blocks. To illustrate the role of risk factors as the dominant predictors Reading Comprehension and Conscientiousness, we provide some example cases. If the average student remained average except for either Reading Comprehension or Conscientiousness (scoring now 2 standard deviations below average on that one predictor, $Z = -2$), their probability of being classified as an inconsistent respondent would be 23 % or 20 %, i.e., $Pr(Y = 1|Z = -2, X = 0) = 1/(1 + exp(-[-2.01 - 2(-0.41)]))$ or $Pr(Y = 1|Z = -2, X = 0) = 1/(1 + exp(-[-2.01 - 2(-0.32)]))$. Similarly, another student scoring equally low ($Z = -2$) on both Conscientiousness and Reading Comprehension (but average score on all other predictors), would be expected to have a 37 % probability of being classified as an inconsistent respondent.

4. Discussion

The finding in the literature that low cognitive abilities are a risk factor for inconsistent responding on a mixed-worded scale (e.g., Bolt et al., 2020; Marsh, 1986; Steedle et al., 2019; Steinmann, Strietholt, & Braeken, 2022) was replicated. Reading comprehension being the dominant factor among other more speed-related or abstract-reasoning measures suggests that inconsistent responding might be more of an interpretative consequence than related to pure perceptual processing.

The finding in the literature that wording-related method factors correlate to personality traits (e.g., Michaelides, Zenger, et al., 2016; Quilty et al., 2006; Schmitt & Stuits, 1985) was clarified from an individual-centered perspective. Among the Big-5 traits, low conscientiousness was the strongest risk factor for inconsistent responding, which is in line with what logically could be expected. To a lesser extent low extraversion and high neuroticism were risk factors for inconsistent responding, which might relate to differences in processing positive and negative wording as found in an electroencephalography study (Ku et al., 2020).

Our finding suggests important roles for both cognitive abilities and personality traits in inconsistent responding. Comparatively speaking, personality traits map mostly to an inattention/diligence mechanism and potentially to an interpretative dimension, whereas cognitive abilities are expected to map to both mechanisms of attention and comprehension difficulty. However, the mechanisms behind inconsistent responses and how these factors map to different cognitive stages of responding remain to be clarified (e.g., Baumgartner, Weijters, & Pieters, 2018).

Note that Steinmann, Strietholt, and Braeken (2022) in one of their illustration examples investigated inconsistent responding among grade 9 students in NEPS, but did not find a similar negative relation between conscientiousness as we do here for the grade 5 students. The difference in findings concerning conscientiousness might be due to a social desirability bias that kicks with older age for these personality self-reports. Regardless, age might be an interesting factor to explore as it might influence both mechanisms of attention and comprehension difficulty.

Further research should assess the generalizability of these findings to other contexts and try to tease out the potential underlying mechanisms. 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). The characteristics of the scales might also have an impact on the response patterns. One may speculate that cognitive abilities play a more crucial role in responding to a mixed-worded scale with more complex wording. Additionally, cultural norms and values may influence the way to interpret mixed-worded items due to different degrees of tolerance for contradiction in different cultures (e.g., Peng &

Nisbett, 1999).

A limitation of the study is related to the personality measures being relatively poor compared with other measures in the study and the fact that the Big-5 scale itself is a mixed-worded scale may have led to correlation artifacts. A sensitivity check was conducted by selecting only PW or NW items of the personality measures and rerunning the analyses. The directions of the correlations between personality traits and class memberships remained unchanged, and the key findings remained robust.

4.1. Conclusion

Although incorporating mixed-worded items has the potential to reduce the tendency toward acquiescent response styles, increase respondents' attentiveness, and provide the opportunity to conduct response consistency checks, the mixed-worded format also risks unintended complications. The link to a person's conscientiousness ironically risks inconsistent responses due to inattentiveness, which might be a realistic concern in a low-stakes context (i.e., inconsistent responses have no repercussions for individuals). The link to reading comprehension and the implied difficulty in processing mixed-worded items raises further caution when using the mixed-worded format with younger kids, second-language learners, or people with reading challenges. Thus, we call for a more cautious and reasoned use of mixed-worded scales seeking the right balance between assessment context, wording complexity, and population characteristics.

CRediT authorship contribution statement

Jianan Chen: Conceptualization, Methodology, Software, Writing – original draft. **Isa Steinmann:** Conceptualization, Writing – review & editing. **Johan Braeken:** Conceptualization, Methodology, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Access to the NEPS data requires the conclusion of a Data Use Agreement with the Leibniz Institute for Educational Trajectories. See https://www.neps-data.de/Data-Center/Data-Access.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.paid.2024.112573.

References

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

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). Edition ZfE* (2nd ed.). Wiesbaden: Springer VS.

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., … 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

Bowling, N. A., Huang, J. L., Bragg, C. B., Khazon, S., Liu, M., & Blackmore, C. E. (2016). Who cares and who is careless? Insufficient effort responding as a reflection of respondent personality. *Journal of Personality and Social Psychology, 111*(2), 218–229. https://doi.org/10.1037/pspp0000085

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://doi.org/10.1027/1015-5759.16.1.13

Gnambs, T., & Schroeders, U. (2020). Cognitive abilities explain wording effects in the Rosenberg self-esteem scale. *Assessment, 27*(2), 404–418. https://doi.org/10.1177/1073191117746503

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology, 27*(1), 99–114. https://doi.org/10.1007/s10869-011-9231-8

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., & Cheung, S. F. (2023). *A constrained factor mixture model for detecting careless responses that is simple to implement*. Organizational Research Methods: Advance online publication. https://doi.org/10.1177/10944281231195298

Ku, L.-C., Chan, S.-h., & Lai, V. T. (2020). Personality traits and emotional word recognition: An ERP study. *Cognitive, Affective, & Behavioral Neuroscience, 20*(2), 371–386. https://doi.org/10.3758/s13415-020-00774-9

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 artifactors? *Journal of Personality and Social Psychology, 70*, 810–819. https://doi.org/10.1037/0022-3514.70.4.810

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., 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 ed.). Los Angeles, CA: Muthén & Muthén.

NEPS Network. (2022). *National Educational Panel Study, scientific use file of starting cohort grade 5. Leibniz Institute for Educational Trajectories (LIfBi), Bamberg*. https://doi.org/10.5157/NEPS:SC3:12.1.0

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.

Peng, K., & Nisbett, R. E. (1999). Culture, dialectics, and reasoning about contradiction. *American Psychologist, 54*(9), 741–754. https://doi.org/10.1037/0003-066X.54.9.741

Quilty, L., Oakman, J., & 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

R Core Team. (2020). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/.

Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton University Press.

Rubin, D. B. (1987). *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. doi: https://doi.org/10.1111/emip.12256.

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

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