To Mix or Not to Mix Positively and Negatively Worded Items

Isa Steinmann, Oslo Metropolitan University, Norway

Preface

This white paper is addressed to questionnaire developers for educational assessment projects, especially those who work with international large-scale assessments. It aims at broadening the perspective on whether to use or abolish mixed-worded scales by drawing upon the empirical study titled "Who Responds Inconsistently to Mixed-worded Scales? Differences by Achievement, Age Group, and Gender" by Isa Steinmann, Jianan Chen, and Johan Braeken, which is published with open access at the journal *Assessment in Education: Principles, Policy & Practice*: https://doi.org/10.1080/0969594X.2024.2318554. The article and white paper were written in course of the IEA Research and Development Fund project "Challenging the Design Principle of Mixed-worded Questionnaire Scales," with Isa Steinmann as principal investigator. I thank Jianan Chen and Johan Braeken for valuable feedback on this white paper.

To Mix or Not to Mix Positively and Negatively Worded Items

Developing good questionnaires can be challenging, especially in the context of international large-scale assessments. In order to obtain reliable and valid data, it should be easy for the respondents to read and understand the item stems and response options, retrieve relevant information from memory, integrate it, and then give the most adequate response (Tourangeau et al., 2000). Usually, multiple items are used in scales to reliably measure target constructs that are not directly observable. In international large-scale assessments, some of the main challenges are to attain a high level of construct coverage and validity across international, multi-language contexts, and this under strict response-time constraints (e.g., Schulz & Carstens, 2020).

The design principle of mixed item wording

One longstanding and widely used questionnaire design principle is to mix both positively and negatively worded items in the same scales in order to prevent respondents from reading and answering merely superficially (e.g., Likert, 1974; Nunnally & Bernstein, 1994). In the Trends in International Mathematics and Science Study (TIMSS) 2019 (TIMSS & PIRLS International Study Center, 2019), for example, fourth-grade students reported on their mathematics self-concept by responding to both positively worded items like "I usually do well in mathematics" alongside negatively worded items like "Mathematics is harder for me than for many of my classmates" (see Figure 1). Students who want to report a positive mathematics self-concept are supposed to agree with positively worded items and disagree with negatively worded ones (see example A in Figure 1). Students who want to report a negative mathematics self-concept are, by contrast, supposed to disagree with positively worded items and agree with negatively worded ones (see example B in Figure 1).

Figure 1. Consistent and inconsistent example responses to the mixed-worded mathematics self-concept scale administered in TIMSS 2019, grade 4

	Consistent responses								Inconsistent responses							
	example A				example B				example C				example D			
	agree a lot	*agree a little*	*disagree a little*	*disagree a lot*	*agree a lot*	*agree a little*	*disagree a little*	*disagree a lot*	*agree a lot*	*agree a little*	*disagree a little*	*disagree a lot*	*agree a lot*	*agree a little*	*disagree a little*	*disagree a lot*
1. I usually do well in mathematics	☒	○	○	○	○	○	☒	○	☒	○	○	○	○	○	☒	○
2. Mathematics is harder for me than for many of my classmates	○	○	○	☒	☒	○	○	○	☒	○	○	○	○	○	○	☒
3. I am just not good at mathematics	○	○	○	☒	○	☒	○	○	○	☒	○	○	○	○	○	☒
4. I learn things quickly in mathematics	☒	○	○	○	○	○	○	☒	☒	○	○	○	○	○	○	☒
5. Mathematics makes me nervous	○	○	☒	○	☒	○	○	○	☒	○	○	○	○	○	☒	○
6. I am good at working out difficult mathematics problems	○	☒	○	○	○	○	○	☒	○	☒	○	○	○	○	○	☒
7. My teacher tells me I am good at mathematics	☒	○	○	○	○	○	☒	○	☒	○	○	○	○	○	☒	○
8. Mathematics is harder for me than any other subject	○	○	☒	○	☒	○	○	○	☒	○	○	○	○	○	☒	○
9. Mathematics makes me confused	○	○	○	☒	☒	○	○	○	☒	○	○	○	○	○	○	☒

Mixed-worded scales are widely used, also in other international large-scale assessments than TIMSS. For example, the OECD states on their webpage that in their Survey on Social and Emotional Skills, "All of the assessment scales use positively and negatively worded items, in view to adjusting for potential response bias" (OECD, 2024).

Inconsistent responses to mixed-worded scales

However, ample empirical evidence suggests that not all respondents give responses that follow the mixed-wording logic of agreeing with one and disagreeing with the other item type. Instead, some respondents are found to either agree (see example C in Figure 1) or disagree (see example D in Figure 1) with both positively and negatively worded items. Using different methods, these empirical studies flagged between 1% and 36% of respondents as answering too similarly to items with opposite meanings in various samples of children, adolescents, and adults from different countries (Arias et al., 2020; Bulut & Bulut, 2022; Chen et al., 2024; García-Batista et al., 2021; Hong et al., 2020; Melnick & Gable, 1990; Steedle et al., 2019; Steinmann et al., 2024; Steinmann, Strietholt, et al., 2022; Steinmann, Sánchez, et al., 2022; Swain et al., 2008). Such respondents can be called inconsistent according to the logic that positively and negatively worded items should evoke opposite answers.

Since inconsistent responses are considered non-meaningful responses, even small shares of respondents answering inconsistently is problematic in itself (e.g., Baumgartner et al., 2018; Steinmann, Strietholt, et al., 2022). Furthermore, the literature suggests that even small proportions of inconsistent respondents affect data quality measures, such as an overestimation of scale dimensionality (e.g., Schmitt & Stults, 1985; Steedle et al., 2019; Steinmann, Sánchez, et al., 2022; Woods, 2006). Some studies also suggest that reliability estimates are impaired in the presence of inconsistent respondents (Arias et al., 2020; Steinmann, Sánchez, et al., 2022). These effects were however not found in all studies (Hong et al., 2020; Steedle et al., 2019).

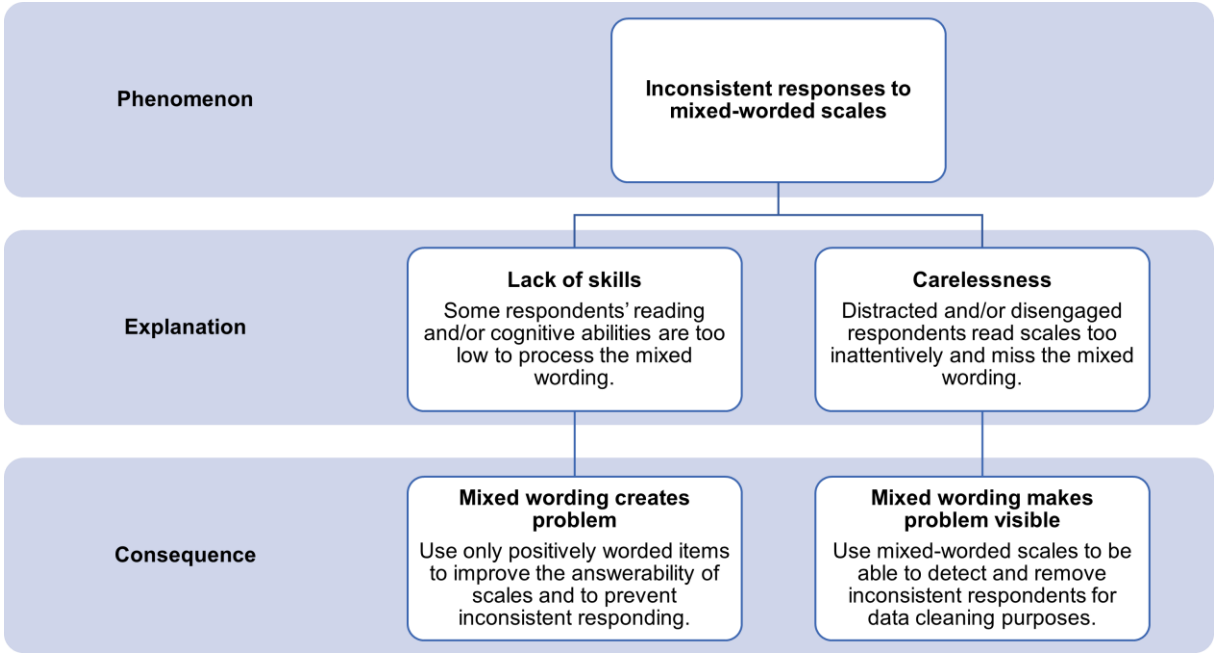
Competing explanations for inconsistent responding

The literature discusses two central explanations for inconsistent responding to mixed-worded scales (see Figure 2), but each has its own consequences and opposing remedies. First, the lack-of-skills explanation assumes that responding to mixed-worded scales is more demanding than answering to scales that contain only positively worded items, and that some respondents lack the required reading and/or cognitive skills to handle the mixed wording successfully (e.g., Baumgartner et al., 2018; Steinmann, Strietholt, et al., 2022; Swain et al., 2008). Such respondents might misread the item stems and response categories with mixed wording and/or have trouble matching their intended response, especially if double negations are required (e.g., "I disagree with being just not good at mathematics"). If this case, the mixed wording would prevent low-performing respondent groups from validly answering the questionnaire. This would not only defy the purpose of survey studies; it would also introduce measurement invariance issues by ability levels. Thus, an unintended effect of mixed-worded scales would be to lead to inconsistent responding, so to speak, *to create a problem*. In view of this lack-of-skills explanation for inconsistent responding, a natural remedy would be to use only positively worded, not mixed-worded scales, to avoid the issue altogether.

Second, the carelessness explanation states that distracted and/or disengaged respondents read scales too superficially to notice and react appropriately to mixed-worded items (e.g., Schmitt & Stults, 1985; Steedle et al., 2019; Steinmann, Strietholt, et al., 2022). Such respondents could, for instance, read only the first item of a scale carefully and then answer all consecutive questions just like the first one. This is a common explanation in the literature, where inconsistent response patterns are often referred to as "careless" or "insufficient effort" responding. If the carelessness explanation is true, using mixed-worded scales would enable researchers to detect and remove such respondents to obtain datasets that contain only respondents who participated diligently in the survey. In other words, mixed wording

would *make a problem visible*. Thus, if this carelessness explanation holds, a logical implication would be to include mixed-worded scales in questionnaires as a data cleaning tool.

Figure 2. Two central explanations for inconsistent responding and their consequences for the use of mixed-worded scales



I would additionally like to mention the related phenomena of acquiescence—"the tendency to choose responses stating agreement regardless of the content of the item" (Primi et al., 2019, p. 2)—and disacquiescence—"the tendency to choose responses stating disagreement regardless of the content of the item" (Primi et al., 2019, p. 2). (Dis-)Acquiescence indexes are often based on the degree of similarity of answers across a series of antonym pairs (e.g., Buchholz, 2022; Primi et al., 2019). In this regard, disacquiescent (cf. example C in Figure 1) and acquiescent (cf. example D in Figure 1) response patterns align with inconsistent responding. However, it is unclear whether (dis-)acquiescence can be assumed to be a third, independent explanation for inconsistent responding, or if it is just another conceptualization for the same phenomenon. Buchholz (2022) for example discusses both a lack of skills and carelessness as potential explanations for acquiescent responding.

Furthermore, I would like to acknowledge that all these explanations rest on the assumption that positively and negatively worded items do indeed work in an opposite way and that inconsistent responses to them are non-meaningful. In other words, it is precluded that giving the same answer to both positively and negatively worded items could convey a coherent, valid statement. It could, however, be discussed that in some cultures and languages, there is more tolerance for contradiction (Peng & Nisbett, 1999), which could imply that not all inconsistent responses are indeed non-meaningful. Also, antonym items might more clearly imply opposite statements than other mixed-worded items of the same scales (e.g., "My teacher tells me I am good at mathematics" and "Mathematics makes me nervous," see Figure 1), which might reflect slightly different underlying constructs.

Empirical evidence on the competing hypotheses

As summarized in Figure 2, there are at least two competing explanations for the same inconsistent-response phenomenon, and these imply opposite consequences for the use of mixed-worded scales. Is it even possible to answer empirically which one applies?

First off, there are numerous studies which have found that respondents who were flagged as inconsistent based on their answers to mixed-worded questionnaire scales performed worse on reading or other scholastic achievement or cognitive ability tests than respondents who gave consistent responses (e.g., Chen et al., 2024; Steedle et al., 2019; Steinmann et al., 2024; Steinmann, Strietholt, et al., 2022; Steinmann, Sánchez, et al., 2022). Steinmann, Chen, et al. (2024) furthermore found that in models that included mathematics achievement, student age, the language spoken at home, and gender as simultaneous predictors of inconsistent responding; mathematics achievement was the strongest predictor across country samples. While at first glance these findings seem to support the lack-of-skills explanation rather than the carelessness explanation, it should be noted that careless respondents might also respond carelessly to tests, not just questionnaires. Thus, upon closer inspection, a negative association between test scores and inconsistent responding can support both competing explanations.

Second, two studies have been conducted to investigate associations between inconsistent responding and proxies for carelessness, namely self-reported conscientiousness (Chen et al., 2024; Steinmann, Strietholt, et al., 2022). While one study found no significant association (Steinmann, Strietholt, et al., 2022), the other one found inconsistent respondents to report lower conscientiousness levels (Chen et al., 2024). Specifically, Chen et al. (2024) regressed inconsistent responding on both ability scores (cognitive reasoning, cognitive speed, reading comprehension, and reading speed) and personality traits (conscientiousness, neuroticism, extraversion, agreeableness, and openness) simultaneously. Low reading comprehension was the strongest predictor of inconsistent responding, followed by low conscientiousness. However, self-reported conscientiousness is not an ideal proxy for carelessness.

Third, studies that try to measure carelessness objectively add valuable evidence to the debate. Baumgartner et al. (2018) conducted an eye-tracking study and found that some respondents' eyes lingered longer at negatively worded items, which would indicate attentiveness, but still gave an inconsistent response. Thus, this finding seems to tentatively support the lack-of-skills explanation over the carelessness explanation. However, the experimental eye-tracking setting might not be comparable with other low-stakes surveys. Future large-scale assessment research could analyze logfile data, for instance, to study associations between inconsistent responding and (too fast) response times as proxies for carelessness.

Fourth, comparing children and adolescents seems to be a promising approach. One can argue that, following the lack-of-skills explanation, children, who are less mature and have lower reading abilities than adolescents, should be more likely to respond inconsistently. Following the carelessness explanation, on the other hand, children could be expected to be more diligent in filling out low-stakes questionnaires than adolescents (Silm et al., 2020) and should thus be less likely to respond inconsistently. Steinmann, Chen, et al. (2024) compared the shares of students who responded inconsistently to a mathematics self-concept scale (see Figure 1) in TIMSS 2019 grades 4 and 8 in all 38 countries that took part in both assessments. They found that the share was significantly larger in grade 4 than in grade 8, which, arguably, supports the lack-of-skills explanation over the carelessness explanation.

Lastly, I would like to point out the large international variation in inconsistent responding. Using data from the joint PIRLS[1]/TIMSS 2011 assessment, Steinmann, Sánchez, et al. (2022) found between 2% of inconsistently responding fourth-graders in Sweden and 36% of inconsistently responding fourth-graders in Honduras. Using data from TIMSS 2019, Steinmann, Chen, et al. (2024) found between 1%

[1] Progress in International Reading Literacy Study

of inconsistently responding eighth-graders in Lithuania and 21% of inconsistently responding fourth-graders in South Africa. While it seems plausible that countries differ somewhat in the carefulness with which the students participate in these low-stakes assessments, this variation seems nonetheless quite large. Furthermore, Steinmann, Chen, et al. (2024) found a strong, negative association between countries' shares of inconsistent respondents and mean achievement levels. This could be interpreted as tentative support for the lack-of-skills explanation, and it suggests that mixed-worded scales might contribute to cross-country measurement invariance issues.

Conclusion: To mix or not to mix positively and negatively worded items?

In conclusion, the empirical literature seems to support both explanations for inconsistent responding: a lack of skills and carelessness (see Figure 2). More research is needed to investigate in which cases (e.g., age groups, low and high stakes conditions) a lack of skills or a lack of carefulness, or possibly both, lead to inconsistent responding, or explore other potential explanations. Both experimental designs and large-scale assessment studies utilizing logfile data seem to be promising avenues for future research. In studies with objective measures for careless responding, it should also be investigated if the assumption holds that fewer respondents answer carelessly to mixed-worded scales than to scales with only positively worded items. Furthermore, future research should investigate the commonalities and differences between the strands of research exploring acquiescence and inconsistent responding.

However, when it comes to the question of using mixed wording in questionnaires or not, I find it important to stress that the recommendation to combine both positively and negatively worded items implicitly rests on the assumption that carelessness is the only explanation for inconsistent responding. The mixed wording aims to prevent too rapid and superficial responding and to enable removing inconsistent respondents for data cleaning purposes. If it is true, however, that at least some respondents—for example young, beginning readers—do not respond carelessly but simply lack the skills to respond to the mixed wording appropriately, mixed wording does not *solve* a problem but *creates* one. Therefore, I recommend using only positively worded items, especially in questionnaires that target populations with low cognitive or reading abilities. Considering the presented research, this holds especially for international large-scale assessments like PIRLS and TIMSS. Other methods to prevent carelessness (e.g., motivate respondents) and detect careless respondents (e.g., too fast response times) are available and might come with fewer risks. If, nevertheless, questionnaire developers decide to use mixed-worded scales to prevent and detect carelessness, I recommend using antonyms that unambiguously express opposite statements.

References

Arias, V. B., Garrido, L. E., Jenaro, C., Martínez-Molina, A., & Arias, B. (2020). A little garbage in, lots of garbage out: Assessing the impact of careless responding in personality survey data. *Behavior Research Methods*, *52*(6), 2489–2505. https://doi.org/10.3758/s13428-020-01401-8

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

Buchholz, J. (2022). *Mixed-worded scales and acquiescence in educational large-scale assessments* (OECD Education Working Papers 269; OECD Education Working Papers, Vol. 269). https://doi.org/10.1787/8dd310c0-en

Bulut, H. C., & Bulut, O. (2022). Item wording effects in self-report measures and reading achievement: Does removing careless respondents help? *Studies in Educational Evaluation*, *72*, 101126. https://doi.org/10.1016/j.stueduc.2022.101126

Chen, J., Steinmann, I., & Braeken, J. (2024). Competing explanations for inconsistent responding to a mixed-worded self-esteem scale: Cognitive abilities or personality? *Personality and Individual Differences, 222,* 112573. https://doi.org/10.1016/j.paid.2024.112573

García-Batista, Z. E., Guerra-Peña, K., Garrido, L. E., Cantisano-Guzmán, L. M., Moretti, L., Cano-Vindel, A., Arias, V. B., & Medrano, L. A. (2021). Using Constrained Factor Mixture Analysis to Validate Mixed-Worded Psychological Scales: The Case of the Rosenberg Self-Esteem Scale in the Dominican Republic. *Frontiers in Psychology*, *12*, 636693. https://doi.org/10.3389/fpsyg.2021.636693

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.

Likert, R. (1974). The method of constructing an attitude scale. In G. M. Maranell (Ed.), *Scaling. A sourcebook for behavioral scientists* (pp. 233–243). Aldine Publ.

Melnick, S. A., & Gable, R. K. (1990). The use of negative item stems. *Educational Research Quarterly*, *14*(3), 31–36.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd edition). McGraw-Hill. http://www.loc.gov/catdir/description/mh022/93022756.html

OECD (2024). *About the OECD's Survey on Social and Emotional Skills*. https://www.oecd.org/education/ceri/social-emotional-skills-study/about/

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

Primi, R., Santos, D., De Fruyt, F., & John, O. P. (2019). Comparison of classical and modern methods for measuring and correcting for acquiescence. *British Journal of Mathematical and Statistical Psychology*, *72*(3), 447–465. https://doi.org/10.1111/bmsp.12168

Schmitt, N., & Stults, D. M. (1985). Factors defined by negatively keyed items: The result of careless respondents? *Applied Psychological Measurement*, *9*(4), 367–373. https://doi.org/10.1177/014662168500900405

Schulz, W., & Carstens, R. (2020). Questionnaire development in international large-scale assessment studies. In H. Wagemaker (Ed.), *Reliability and Validity of International Large-Scale Assessment* (Vol. 10, pp. 61–83). Springer International Publishing. https://doi.org/10.1007/978-3-030-53081-5_5

Silm, G., Pedaste, M., & Täht, K. (2020). The relationship between performance and test-taking effort when measured with self-report or time-based instruments: A meta-analytic review. *Educational Research Review*, *31*, 100335. https://doi.org/10.1016/j.edurev.2020.100335

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

Steinmann, I., Chen, J., & Braeken, J. (2024). Who Responds Inconsistently to Mixed-Worded Scales? Differences by Achievement, Age Group, and Gender. *Assessment in Education: Principles, Policy & Practice,* 1–28. https://doi.org/10.1080/0969594X.2024.2318554

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

Swain, S. D., Weathers, D., & Niedrich, R. W. (2008). Assessing Three Sources of Misresponse to Reversed Likert Items. *Journal of Marketing Research*, *45*(1), 116–131. https://doi.org/10.1509/jmkr.45.1.116

TIMSS & PIRLS International Study Center (2019). *TIMSS: Trends in International Mathematics And Science Study*. https://timssandpirls.bc.edu/timss-landing.html

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The Psychology of Survey Response* (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511819322

Woods, C. M. (2006). Careless Responding to Reverse-Worded Items: Implications for Confirmatory Factor Analysis. *Journal of Psychopathology and Behavioral Assessment*, *28*(3), 186–191. https://doi.org/10.1007/s10862-005-9004-7