Integrated Science Education Journal
Vol. 7, No. 3, May 2026, pp. 459~469
ISSN: 2716-3725, DOI: 10.37251/isej.v7i3.2960

Research Article

Negative-Worded Items Functioning as Method Artifacts in the Chemistry Identity Scale: Evidence from Exploratory, Confirmatory, and Bifactor Analyses

Yuleks Juru Mudi[1], Aeda Kasrianti[1], Sefthy P. B. Syahailatua[1], Nurul Isnaini[1], Balthasar Eba[1]
[1] Study Program of Educational Research and Evaluation, Yogyakarta State University, Yogyakarta, Indonesia

Article Info

Article history:

Received Mar 28, 2026
Revised Apr 29, 2026
Accepted May 23, 2026
OnlineFirst May 30, 2026

Keywords:

Chemistry Identity
Method Bias
Negative-Worded Items
Psychometric Modeling
Self-Report Measurement

ABSTRACT

Purpose of the study: Chemistry identity is an important affective construct in science education because it is associated with learning engagement, academic persistence, and STEM career aspirations. This study aims to evaluate whether negatively worded items represent substantive dimensions of the construct or merely methodological artifacts.

Methodology: This study involved 300 senior high school students in Indonesia who completed the Chemistry Identity Scale, consisting of 27 items, including five negatively worded items. Data were analyzed using a comprehensive psychometric approach that incorporated exploratory factor analysis (EFA), confirmatory factor analysis (CFA), and bifactor modeling to distinguish substantive construct variance from method variance attributable to item wording.

Main Findings: The findings showed that negatively worded items tended to form a distinct cluster during the exploratory stage, indicating shared method variance. The best-fitting CFA model was the four-factor model with an additional negative wording method factor. Bifactor analysis revealed the dominance of a general chemistry identity factor; however, negatively worded items contributed minimally to the general construct, suggesting that these items function more as sources of method variance than as substantive indicators.

Novelty/Originality of this study: The novelty of this study lies in its comprehensive evaluation of wording effects in chemistry identity measurement through the integration of EFA, competitive CFA, and bifactor modeling. These findings have practical implications for educational instrument developers, highlighting the need for greater caution when using negatively worded items, as they may affect score interpretation and lead to less accurate evaluative decisions.

Corresponding Author:
Yuleks Juru Mudi,
Study Program of Educational Research and Evaluation, Graduate School, Universitas Negeri Yogyakarta,
Jl. Colombo No. 1, Karangmalang, Yogyakarta, 55281, Indonesia
Email: yuleksmudi2000@gmail.com

1. INTRODUCTION

In science education, chemistry identity has increasingly been recognized as an important affective construct because it reflects how students position themselves in relation to chemistry learning. This construct generally encompasses dimensions such as recognition, competence/performance, interest, and sense of

belonging [1]-[3]. A growing body of research has shown that chemistry identity is associated with students' engagement in learning, academic persistence, and even career aspirations in STEM fields [1], [3], [4]. Therefore, the quality of instruments used to measure chemistry identity is critically important. The scores generated should accurately reflect the intended construct rather than being influenced by the technical characteristics of the instrument itself.

One of the most frequently discussed methodological issues in the psychometric literature is the use of negatively worded items in self-report instruments. This strategy was originally intended to reduce acquiescence bias, namely the tendency of respondents to agree with statements without carefully considering their content. However, empirical evidence has shown that the inclusion of negatively worded items may instead introduce method variance, that is, systematic variance attributable to item format rather than the substantive construct being measured [5]-[8]. The consequences are far from trivial. Factor structures may become distorted, reliability may decrease, and score interpretations may become less accurate [5], [9], [10]. For example, [9] demonstrated that scales combining positively and negatively worded items tend to exhibit lower reliability and less stable factor structures than scales with consistent item wording. Similarly, [7] found that different item reversal strategies can lead to non-equivalent latent structures.

This issue becomes even more evident when examined from a response process perspective. [11] demonstrated that negatively worded items impose a greater cognitive burden, as reflected in longer processing times and a greater need for reinterpretation compared with positively worded items. [12] further emphasized that wording effects are not solely related to cognitive complexity, but also involve linguistic factors that influence how respondents interpret and reason about item meaning. In a more applied context, [13] showed that negative wording alone can produce inconsistent response patterns and even affect model fit. In other words, factors emerging from negatively worded items do not necessarily reflect substantively meaningful psychological dimensions, but may instead represent methodological artifacts.

In science education, science identity and chemistry identity instruments have been widely used to understand the development of students' academic identities as well as to evaluate their affective engagement in learning [1], [2], [14]. However, most studies in this area have primarily focused on the conceptual development of the construct, the process of identity formation, or the relationship between identity and other learning-related variables [2], [3], [15]. Psychometric evaluations that specifically examine the consequences of item design, particularly in relation to negative wording, remain relatively limited. Yet, within the broader fields of psychological and educational measurement, substantial evidence has shown that mixing positively and negatively worded items can produce artificial factor structures that do not fully reflect the intended construct [13], [16], [17].

This is where the research gap underlying the present study emerges. Although the effects of negative wording have been extensively investigated across various psychological scales, empirical evidence specifically examining its impact on chemistry identity measurement remains limited [2], [18]. In addition, most previous studies have tended to rely on a single analytical approach, making it difficult to comprehensively disentangle substantive construct variance from method variance [19]. This gap is particularly important because chemistry identity scores are frequently used as a basis for drawing conclusions about students' affective engagement in chemistry learning. If the factor structure of the instrument is influenced more by item wording than by the construct it is intended to measure, interpretations of those scores may be misleading.

Based on these considerations, this study evaluates the psychometric structure of the Chemistry Identity Scale, with particular emphasis on the impact of negatively worded items. Unlike previous studies that have relied on a single analytical approach, the present study integrates exploratory factor analysis (EFA), confirmatory factor analysis (CFA), and bifactor modeling to examine whether the factors emerging from negatively worded items genuinely represent meaningful latent dimensions or are more appropriately interpreted as methodological artifacts [19]-[21]. This approach enables a more comprehensive evaluation of the distinction between substantive construct variance and method variance, thereby providing a stronger foundation for interpreting scores derived from self-report measures of affective constructs.

Based on the foregoing rationale, this study aims to examine the presence and magnitude of method variance associated with negatively worded items in the measurement of chemistry identity. Specifically, the study is guided by the following research questions:

1. How do negatively worded items behave within the exploratory factor structure of the Chemistry Identity Scale?
2. Do negatively worded items represent substantive dimensions or wording-related method factors in confirmatory models?
3. To what extent does bifactor modeling support the dominance of a general chemistry identity factor relative to wording-related variance?

By addressing these questions, this study is expected to contribute to the development of chemistry identity instruments that are more valid, reliable, and accurately interpretable, while also enriching the methodological discourse on the evaluation of self-report instruments in science education.

2. RESEARCH METHOD

This study employed a psychometric validation design using a cross-sectional survey approach to evaluate whether negatively worded items in the Chemistry Identity Scale represent substantive dimensions of chemistry identity or instead introduce method variance that may affect score interpretation. The primary focus of this study was to evaluate the quality of the measurement structure of a self-report instrument rather than to examine causal relationships or the effectiveness of an intervention.

The participants consisted of 300 senior high school students (Grades 10–12) from various regions of Indonesia who were enrolled in chemistry courses at the time of data collection. This sample size was considered adequate for factor analysis, given that sample size sufficiency in factor analytic studies depends on model complexity, the number of indicators, and parameter characteristics, with a moderate sample size such as 300 generally regarded as sufficient for well-defined models [22], [23]. Participants were recruited using a convenience sampling technique through an open online survey. This sampling approach was deemed appropriate given that the study focused on evaluating the psychometric characteristics of the instrument, particularly item behavior and construct validity, rather than on broad population generalization. All participation was voluntary and anonymous.

The instrument used in this study was the Chemistry Identity Scale, adapted from previously developed science identity and STEM identity instruments [14], [24]. The instrument measures four theoretical dimensions, namely recognition, performance/competence, interest, and sense of belonging, which represent the social, affective, and cognitive aspects of students' identity formation within the context of chemistry learning. In total, the instrument consisted of 27 items rated on a four-point Likert scale ranging from 1 (strongly disagree) to 4 (strongly agree), including five negatively worded items (Items 2, 7, 10, 21, and 27). These items were originally included to reduce acquiescence bias, defined as the tendency of respondents to agree with statements without carefully considering the item content [25]. However, a growing body of psychometric research suggests that negatively worded items may instead introduce wording effects or method effects that can distort factor structures and reduce measurement quality [6], [26].

Before administration, the instrument was reviewed through expert judgment by four specialists in educational psychology, chemistry education, and educational measurement and evaluation. This review aimed to assess content relevance, construct clarity, and linguistic appropriateness. Feedback from the experts was used to refine the instrument prior to data collection. Data were collected through Google Forms between October and November 2025, and all participants provided informed consent before taking part in the study.

Data analysis was conducted using R software. Because the data were obtained from a four-point Likert scale and were therefore ordinal in nature, polychoric correlations were used, as they are considered more appropriate than conventional Pearson correlations for estimating relationships among latent variables in ordinal data [27]. Prior to conducting factor analysis, data suitability was assessed using the Kaiser–Meyer–Olkin (KMO) measure and Bartlett's test of sphericity to evaluate sampling adequacy and the factorability of the correlation matrix.

The number of empirically supported factors was determined using parallel analysis, which is widely recommended as one of the most robust approaches for factor retention, including for ordinal data [28]. Subsequently, exploratory factor analysis (EFA) was conducted using principal axis factoring with oblimin rotation to explore the latent structure and identify the possible emergence of additional factors associated with negatively worded items. This approach was selected because it is appropriate for evaluating latent constructs based on common variance and allows for correlations among latent factors, which are conceptually plausible in this context [29], [30].

To provide a more rigorous test of the measurement structure, the analysis was extended using confirmatory factor analysis (CFA) implemented in the lavaan package with the Weighted Least Squares Mean and Variance Adjusted (WLSMV) estimator, which is recommended for ordinal indicators because it provides more appropriate parameter estimates than approaches that assume continuous data [31]. In this study, five competing measurement models were compared: a one-factor model, the theoretical four-factor model, a positive–negative wording model, a four-factor model with an additional negative method factor, and a bifactor model. Model fit was evaluated using CFI, TLI, RMSEA, and SRMR, following current methodological recommendations [32], [33].

To gain a deeper understanding of the contributions of substantive variance and method variance, the analysis was further extended using bifactor modeling. Several additional indices were calculated, including Explained Common Variance (ECV), Omega Hierarchical (ωH), Percent of Uncontaminated Correlations (PUC), and Item-level Explained Common Variance (IECV), which were used to evaluate the dominance of the general factor as well as the contribution of specific factors within the measurement model [19], [24], [34]. This study was conducted in accordance with ethical principles for research involving human participants. All participants were informed about the purpose of the study, data confidentiality, and their right to withdraw from participation at any time without consequence. All responses were collected anonymously and used solely for academic purposes.

3. RESULTS AND DISCUSSION

3.1. Exploratory Evidence of Negative Item Behavior

Before evaluating the behavior of negatively worded items within the factor structure of the Chemistry Identity Scale, internal reliability and data suitability were first assessed to ensure that the instrument provided an adequate psychometric foundation for subsequent analyses. The results indicated that the instrument demonstrated good overall internal consistency, with a Cronbach's alpha coefficient of 0.878, suggesting a reasonably consistent pattern of inter-item responses in representing the intended construct. However, Cronbach's alpha should be interpreted with caution, as it does not automatically guarantee the structural validity of an instrument, particularly when the construct is multidimensional in nature [18], [35]. At the dimensional level, the reliability coefficients showed some variation, with performance/competence demonstrating relatively higher reliability ($\alpha = 0.730$), whereas recognition ($\alpha = 0.637$), interest ($\alpha = 0.654$), and sense of belonging ($\alpha = 0.671$) exhibited comparatively lower internal consistency. This variation may indicate the presence of additional sources of variance influencing response patterns, including the possible influence of method effects associated with item wording characteristics [19], [36].

Data suitability for factor analysis was also examined using the Kaiser–Meyer–Olkin (KMO) measure and Bartlett's Test of Sphericity, which are commonly used to assess sampling adequacy and data factorability prior to conducting factor analysis [37]-[39]. The KMO value of 0.88 indicated excellent sampling adequacy, while the significant Bartlett's test result ($p < 0.001$) suggested that the inter-item correlations were sufficient for factor analysis [40]. Overall, these findings indicate that the data were suitable for further analysis using both exploratory and confirmatory approaches. To address the first research question concerning the behavior of negatively worded items within the exploratory factor structure, a parallel analysis based on polychoric correlations was conducted to determine the number of factors empirically supported by the data before further exploration of the latent structure. This approach was selected because the data were derived from an ordinal Likert scale, making polychoric correlations more appropriate for representing inter-item relationships more accurately than conventional Pearson correlations [27], [41].

Figure 1 presents the results of the parallel analysis based on a comparison between the empirical eigenvalues and those generated from simulated data. Visually, the eigenvalues from the observed data remained above the simulated values beyond the theoretical four-dimensional structure underlying the Chemistry Identity Scale. This finding suggests that the empirical data support a more complex structure than the initial theoretical model, indicating the presence of additional sources of variance beyond the four core dimensions of chemistry identity that were conceptually intended to be measured. This phenomenon is consistent with previous findings showing that rating scale data may exhibit a tendency toward overfactoring when additional non-substantive sources of variance are present in response patterns [42].

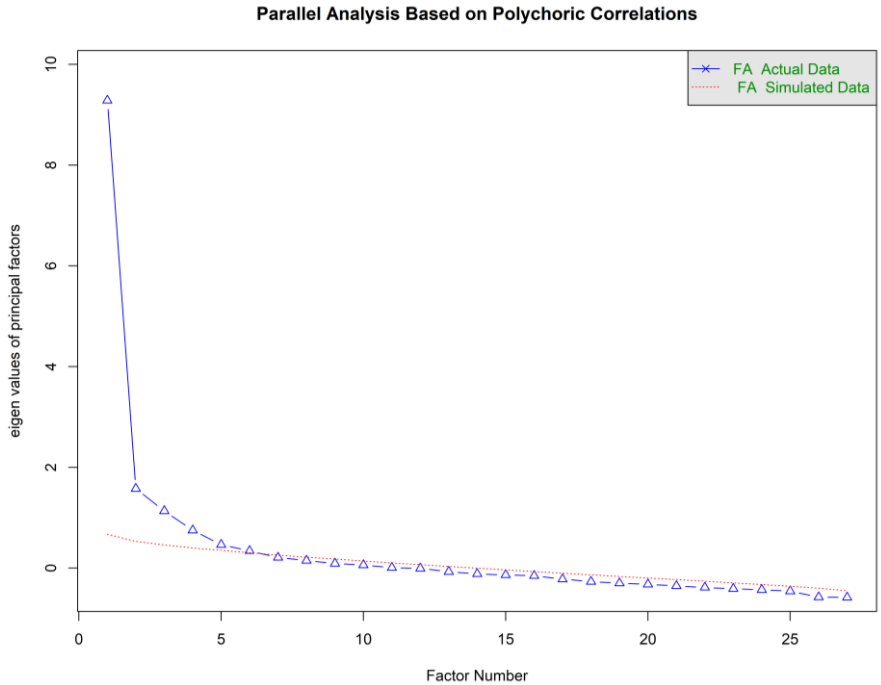

Figure 1. Parallel analysis based on polychoric correlations for factor retention in the Chemistry Identity Scale.

These preliminary findings were further clarified through exploratory factor analysis (EFA), which revealed that negatively worded items tended to form a distinct cluster, separate from the primary theoretical

dimensional structure. This pattern suggests that negatively worded items do not behave entirely equivalently to positively worded items in representing the chemistry identity construct. When items that are conceptually intended to measure the same construct instead cluster according to their wording format, this indicates that respondents' answers may be influenced not only by their level of chemistry identity, but also by the technical characteristics of the items themselves.

This interpretation is consistent with findings from the contemporary psychometric literature showing that negatively worded items often generate additional factor structures that do not necessarily reflect substantively meaningful psychological dimensions. [5] demonstrated that negatively worded items can produce substantial systematic variance that affects the overall measurement structure. Similarly, [9] found that combining positively and negatively worded items within a single instrument can weaken dimensional stability and reduce the psychometric quality of the measure. Likewise, [7] showed that the inclusion of reverse-worded items can make an instrument that is theoretically unidimensional appear multidimensional, as positively and negatively worded items tend to form separate factors.

Similar findings were reported by [13], who demonstrated that separating items based on positive versus reverse wording significantly improved model fit, suggesting that the additional factor structure observed was more likely attributable to wording effects than to a genuinely new substantive construct. [6] further showed that this phenomenon becomes even more pronounced among respondents with lower cognitive processing capacity, indicating that wording effects are not merely a statistical concern, but are also closely related to respondents' response processes.

Nevertheless, these exploratory findings should be interpreted with caution. The emerging factor structure cannot be evaluated solely on the basis of statistical separation among items; it is equally important to consider whether such groupings are conceptually meaningful in representing the intended construct [43]. In this context, the emergence of a cluster of negatively worded items cannot be immediately regarded as merely a methodological artifact. [5] cautioned that factors formed by negatively worded items do not always reflect only noise or measurement error, but may, under certain conditions, also contain substantive meaning. Therefore, at this stage, the EFA findings are more appropriately interpreted as preliminary evidence of possible wording-related variance rather than as a definitive conclusion regarding the nature of the observed structure. These findings subsequently provided the basis for further examination through confirmatory analysis.

3.2. Confirmatory Evidence of Wording Effects

To address the second research question, namely whether negatively worded items represent distinct substantive dimensions or instead form wording-related method factors, confirmatory factor analysis (CFA) was conducted by comparing five competing measurement models. These models included a one-factor model, the theoretical four-factor model, a positive–negative wording model, a four-factor model with an additional negative method factor, and a bifactor model.

Table 1. Comparison of Measurement Model Fit Indices

Model	CFI	TLI	RMSEA	SRMR
1 Factor	0.941405771	0.936522919	0.110890038	0.09692834
4 Factor	0.951846307	0.946849226	0.101470279	0.092457163
Positive-Negative	0.957976403	0.95433349	0.094055234	0.085246314
4 Factor + Negative Method	0.970389263	0.96679435	0.080202897	0.078128473
Bifactor	0.959808799	0.95577708	0.092556679	0.083472085

The analysis showed that the four-factor model with an additional negative method factor provided the best fit compared with the other competing models. This finding conveys an important implication. On the one hand, the theoretical structure of the Chemistry Identity Scale, consisting of recognition, performance/competence, interest, and sense of belonging, continued to receive empirical support. On the other hand, the relationships among items could not be fully explained solely by the chemistry identity construct itself, as additional variance was systematically associated with the negatively worded items.

The one-factor model demonstrated the poorest fit, indicating that chemistry identity cannot be adequately reduced to a single, fully homogeneous construct. This finding is consistent with the conceptual framework underlying the Chemistry Identity Scale, which conceptualizes chemistry identity as a multidimensional construct [2], [3]. The theoretical four-factor model showed improved fit, confirming that the instrument's foundational conceptual structure remains empirically relevant.

However, the most substantial improvement in model fit occurred when a specific method factor for negatively worded items was incorporated into the model. Figure 2 illustrates this best-fitting model, in which the items continued to load onto their respective substantive dimensions, while the negatively worded items also shared additional variance through a dedicated method factor (negative wording method factor). Visually, this

model suggests that negatively worded items function not only as indicators of the chemistry identity construct, but also exhibit systematic associations arising from their shared wording format.

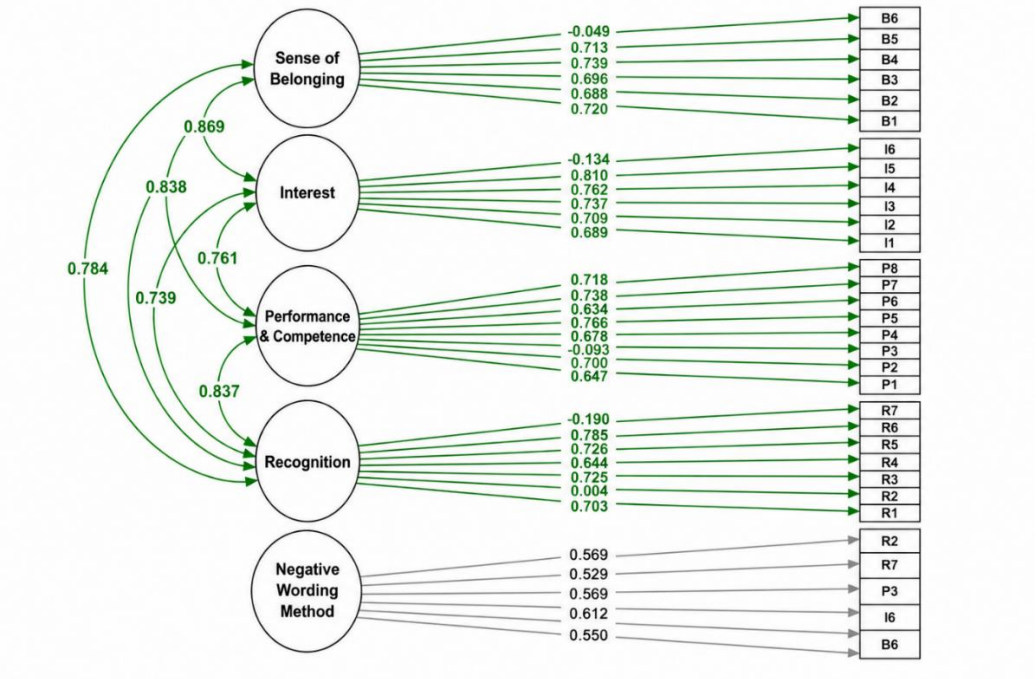

Figure 2. Best-fitting confirmatory factor model showing the four substantive dimensions and the negative wording method factor.

These findings suggest that some of the covariance among negatively worded items cannot be fully accounted for by the substantive chemistry identity dimensions alone, but instead reflect an additional source of systematic variance associated with item wording characteristics. In other words, the negatively worded items appear to behave differently from the positively worded indicators, contributing variance that is not entirely attributable to the intended substantive construct.

This interpretation is further supported by the standardized factor loadings shown in Figure 2. Several negatively worded items displayed negative loadings on their substantive factors, including R7 (-0.190), P3 (-0.093), I6 (-0.134), and B6 (-0.049), while simultaneously loading moderately on the negative wording method factor. This pattern indicates that these items may function differently from the positively worded indicators, reinforcing the interpretation that wording-related method variance is present in the measurement model.

This interpretation is consistent with findings from previous psychometric studies. [13] demonstrated that when items were separated based on positive versus reverse wording, model fit improved significantly, suggesting that wording itself may represent a distinct source of systematic variance. Similarly, [6] found that models explicitly accounting for wording effects provided a better representation than simple unidimensional models, particularly when negatively worded items were included. Comparable findings were reported by [7], who showed that reverse-worded items may introduce an additional method-related latent dimension beyond the intended substantive dimensions, reflecting item format effects rather than the construct itself. Consistent with this interpretation, [44] also demonstrated that the direction of item wording can generate systematic method effects that influence the structure of measurement models.

Further support comes from [10], who reported that combining positively and negatively worded items within a single model can reduce model fit quality, whereas models that explicitly separate wording effects often perform better. More broadly, [9] noted that mixed item wording may weaken the structural validity of an instrument, even when overall scale scores appear relatively stable.

Nevertheless, these findings should be interpreted with caution. As emphasized by [5], the emergence of a factor associated with negatively worded items does not necessarily mean that the factor is purely a methodological artifact. In some contexts, such factors may still contain a degree of substantive variance. Therefore, the negative wording method factor identified in this study is more appropriately interpreted as evidence of substantial wording-related variance, rather than definitive proof that the variance is entirely methodological.

From a methodological perspective, these findings also highlight that good statistical model fit alone does not automatically indicate that the observed latent structure fully reflects the intended psychological construct. As [19] cautioned, the evaluation of measurement models should not rely solely on fit indices, but

should also consider whether the sources of variance shaping the model are substantively meaningful or partly influenced by measurement artifacts. In this context, the CFA results suggest that the observed structure of chemistry identity is influenced, at least in part, by wording effects, and therefore interpretation of the construct should be undertaken with appropriate caution.

3.3. Bifactor Evidence for General Chemistry Identity

To address the third research question, namely the extent to which bifactor modeling supports the dominance of a general chemistry identity factor relative to wording-related variance, the analysis was extended using a bifactor model. This approach was selected because it allows for a clearer separation between variance attributable to the primary substantive construct and variance arising from specific factors or other non-substantive sources. Unlike second-order models, which represent the influence of a general factor indirectly through first-order factors, bifactor models allow each item to load directly onto both the general factor and its corresponding specific factor, thereby providing a more detailed evaluation of the sources of measurement variance [19], [36], [45].

The analysis indicated that the general chemistry identity factor remained highly dominant, with an Explained Common Variance (ECV) of 0.859, an Omega Hierarchical of 0.858, and a Percent of Uncontaminated Correlations (PUC) of 0.972. Based on established guidelines for interpreting bifactor models, this combination of values suggests that the majority of shared variance among the items is still explained by a single strong general construct, namely chemistry identity [19]. These findings indicate that although some response variation is associated with item wording, the instrument's primary substantive structure remains dominated by the psychologically meaningful construct of chemistry identity.

These findings provide an important clarification of the earlier CFA results. In the CFA stage, the model including an additional negative method factor demonstrated the best fit, indicating the presence of wording effects within the measurement structure. However, the bifactor results suggest that the influence of this method effect is not sufficiently strong to override the dominance of the general chemistry identity factor. In other words, although wording effects are present as an additional source of variance, they do not completely obscure the presence of the primary substantive construct being measured. This interpretation is consistent with the modern psychometric literature, which emphasizes that better model fit does not necessarily imply that the specific factors that emerge possess strong substantive meaning. [19] argued that the evaluation of bifactor models should not rely solely on goodness-of-fit indices, but must also consider whether the general factor genuinely dominates the measurement structure. Findings from [36], and [45] likewise suggest that constructs that appear theoretically multidimensional may still be empirically dominated by a single strong general factor, requiring cautious interpretation of the subfactors.

Nevertheless, when item contributions were examined in greater detail through Item-level Explained Common Variance (IECV), a more complex pattern emerged, as illustrated in Figure 3. The distribution of IECV values revealed a clear contrast between positively and negatively worded items. Positively worded items demonstrated relatively high and consistent contributions to the general factor, whereas negatively worded items exhibited substantially lower contributions, with some approaching zero. This pattern suggests that not all items contributed equally to the measurement of the general chemistry identity construct. More specifically, the negatively worded items appeared to make only a limited contribution to the primary construct, despite formally remaining part of the same instrument.

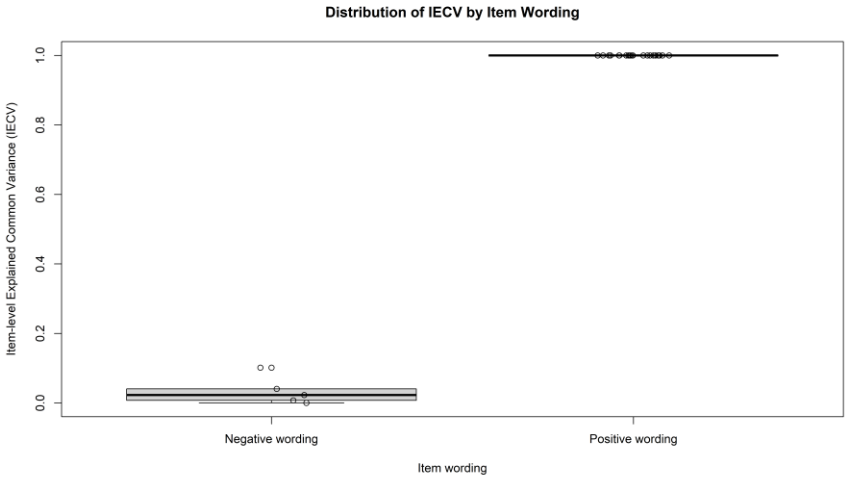

Figure 3. Distribution of item-level explained common variance (IECV) comparing positively and negatively worded items.

From a substantive perspective, these findings suggest that responses to negatively worded items are likely not driven entirely by chemistry identity itself, but also by additional response processes associated with the way the items are formulated. In other words, some negatively worded items appear to be more sensitive to the format in which the statements are presented than to the substantive content of the construct they are intended to measure.

This interpretation is supported by the response process literature. Using an eye-tracking approach, [11] demonstrated that negatively worded items require longer reading times, extended response times, and higher frequencies of response revision compared with positively worded items. These findings suggest that negatively worded items impose additional cognitive processing demands during the item comprehension stage. [12] further emphasized that negative wording can alter the way respondents linguistically interpret item content, meaning that the resulting responses may not always purely reflect the psychological construct being measured.

On the other hand, [46] offers an important interpretive nuance by showing that factor separation among negatively worded items does not necessarily arise solely from respondent confusion, but may also be influenced by the semantic characteristics of the items, such as extreme wording or specific linguistic structures. Therefore, the findings of the present study do not necessarily imply that all negatively worded items are inherently problematic; rather, they suggest that such items may introduce additional sources of variance that are not identical to the primary substantive construct.

In the educational context, these findings carry significant implications. Chemistry identity instruments are frequently used to assess students' affective engagement, the development of academic identity, and even tendencies toward persistence in STEM fields. If some items capture responses to wording more strongly than the intended construct, interpretations of students' scores may become less accurate. Therefore, although the overall instrument score remains supported by the dominance of a strong general factor, the quality of each item's contribution should still be critically considered.

Overall, the bifactor modeling results reinforce the interpretation that the Chemistry Identity Scale retains a strong general construct foundation, while negatively worded items demonstrate psychometrically less optimal contributions to the measurement of that general factor. Accordingly, the wording effect identified in this study is more appropriately understood as a source of method-related variance affecting measurement quality at the item level, rather than as evidence that the chemistry identity construct itself has lost substantive validity.

The findings of this study provide important implications, both theoretically and practically, particularly for the development of self-report affective instruments in the context of science education. From a theoretical perspective, the results reinforce the argument that the emergence of additional factors associated with negatively worded items should not be automatically interpreted as substantively meaningful new psychological dimensions. The combined evidence from exploratory factor analysis, confirmatory factor analysis, and bifactor modeling suggests that some of the additional structure observed is more appropriately understood as method-related variance rather than as the representation of a new latent construct. These findings are consistent with the psychometric literature showing that reverse-worded items often produce artificial multidimensionality that reflects method effects more than substantive constructs [6], [7], [13].

This study also makes a methodological contribution to the chemistry identity literature, which has largely focused on the relationships between identity, learning engagement, academic persistence, and STEM career aspirations [1]-[3], while remaining relatively limited in evaluating how item design may affect measurement validity. In this way, the present study extends the discussion beyond substantive issues toward the methodological evaluation of instrument quality. The primary novelty of this study lies in the use of a comprehensive evaluative approach that integrates exploratory factor analysis, competitive confirmatory factor analysis, and bifactor modeling to examine whether the observed factors genuinely represent substantive constructs or are merely methodological artifacts.

From a practical perspective, these findings have important implications for educational instrument developers, particularly in encouraging greater caution in the use of negatively worded items as an instrument design strategy. This approach has long been employed as a means of reducing acquiescence bias, although its effectiveness remains debated within the psychometric literature [26], [46]. However, the findings of the present study suggest that the inclusion of negatively worded items may instead introduce wording-related method variance that can affect score interpretation. In the context of educational evaluation, this implication is particularly important because chemistry identity instruments are often used to inform decisions regarding students' affective engagement, instructional design, and the development of STEM education programs. If part of the scores is influenced by wording artifacts, educational decisions based on such measurement results may become less accurate.

Nevertheless, several limitations should be considered when interpreting the findings of this study. First, the exploratory and confirmatory analyses were conducted using the same sample, which limits the strength of cross-validation, as the identified structure has not yet been tested on an independent sample [47]. Second, the use of convenience sampling restricts the generalizability of the findings to broader student

populations [48], [49]. Third, this study was conducted within the Indonesian student context, meaning that the potential influence of linguistic and cultural characteristics on the emergence of wording effects cannot be fully disentangled, particularly given that survey response styles may also be shaped by cultural factors at the individual level [11], [12], [50]. Fourth, this study focused on a single specific instrument, namely the Chemistry Identity Scale, and therefore caution is warranted in extending these conclusions to other affective instruments in STEM education that may differ in construct characteristics and item design.

Based on these limitations, future research is recommended to use independent samples for the exploratory and confirmatory stages, conduct measurement invariance testing across different groups, and consider alternative approaches such as Item Response Theory (IRT) or response process-based studies to gain a deeper understanding of the cognitive mechanisms underlying wording effects [51], [52]. In addition, it is important to examine whether similar patterns also emerge in other affective instruments within STEM education, so that a more comprehensive understanding of the impact of wording on measurement validity can be established.

4. CONCLUSION

This study demonstrates that negatively worded items in the Chemistry Identity Scale do not function entirely equivalently to positively worded items in representing the chemistry identity construct. At the exploratory stage, negatively worded items tended to form a distinct cluster, indicating the presence of additional sources of variance beyond the primary theoretical structure. Subsequent confirmatory analysis showed that the model including a negative method factor provided the best fit, suggesting that part of the relationships among negatively worded items was more closely associated with wording characteristics than solely with the chemistry identity construct being measured.

Nevertheless, the bifactor modeling results indicate that chemistry identity remains dominated by a strong general factor. These findings suggest that the presence of wording effects does not completely undermine the substantive validity of the instrument, but rather affects the quality of contribution of certain items, particularly negatively worded ones. Thus, negatively worded items in affective instruments do not necessarily represent new psychological dimensions, but may instead function as sources of method-related variance that should be explicitly considered in construct validity evaluation.

From a theoretical perspective, this study extends the chemistry identity literature by demonstrating that construct validity is not solely concerned with the adequacy of theoretical model fit, but also with the ability to distinguish between substantive variance and variance arising from instrument design. From a practical perspective, these findings have important implications for instrument developers and educational evaluators, encouraging greater caution in the use of negatively worded items as an instrument design strategy, as wording effects may influence score interpretation and the quality of educational decisions based on measurement results if not adequately evaluated.

ACKNOWLEDGEMENTS

The authors would like to express their sincere gratitude to all students who participated in this study and contributed valuable data to this research. Appreciation is also extended to all parties who supported the data collection process. This study was conducted without any external funding.

AUTHOR CONTRIBUTIONS

Conceptualization, Yuleks Juru Mudi and Aeda Kasrianti; Methodology, Yuleks Juru Mudi and Aeda Kasrianti; Data curation, Yuleks Juru Mudi; Formal analysis, Yuleks Juru Mudi and Sefthy P. B. Syahailatua; Investigation, Yuleks Juru Mudi; Writing – original draft preparation, Yuleks Juru Mudi; Writing – review and editing, Aeda Kasrianti, Sefthy P. B. Syahailatua, Nurul Isnaini, and Balthasar Eba; Supervision, Aeda Kasrianti and Balthasar Eba. All authors have read and agreed to the published version of the manuscript.

CONFLICTS OF INTEREST

The author(s) declare no conflict of interest.

USE OF ARTIFICIAL INTELLIGENCE (AI)-ASSISTED TECHNOLOGY

The authors declare that no artificial intelligence (AI) tools were used in the generation, analysis, or writing of this manuscript. All aspects of the research, including data collection, interpretation, and manuscript preparation, were carried out entirely by the authors without the assistance of AI-based technologies.

REFERENCES

[1] X. Guo, W. Deng, K. Hu, W. Lei, S. Xiang, and W. Hu, "The effect of metacognition on students' chemistry identity: the chain mediating role of chemistry learning burnout and chemistry learning flow," *Chem. Educ. Res. Pract*., vol. 23,

no. 2, pp. 408–421, 2022, doi: 10.1039/D1RP00342A.
[2] K. N. Hosbein and J. Barbera, "Development and evaluation of novel science and chemistry identity measures," *Chem. Educ. Res. Pract*., vol. 21, no. 3, pp. 852–877, 2020, doi: 10.1039/C9RP00223E.
[3] Z. Jiang, B. Wei, S. Chen, and L. Tan, "Examining the formation of high school students' science identity," *Sci. Educ.,* vol. 33, no. 1, pp. 135–157, Feb. 2024, doi: 10.1007/s11191-022-00388-2.
[4] Z. Hazari, G. Sonnert, P. M. Sadler, and M.-C. Shanahan, "Connecting high school physics experiences, outcome expectations, physics identity, and physics career choice: A gender study," *J. Res. Sci. Teach.,* vol. 47, no. 8, pp. 978–1003, 2010, doi: 10.1002/tea.20363.
[5] V. B. Arias and B. Arias, "The negative wording factor of Core Self-Evaluations Scale (CSES): Methodological artifact, or substantive speci fi c variance ?," *Pers. Individ. Dif*., vol. 109, pp. 28–34, 2017, doi: 10.1016/j.paid.2016.12.038.
[6] H. C. Bulut and O. Bulut, "Item wording effects in self-report measures and reading achievement: Does removing careless respondents help?," *Stud. Educ. Eval*., vol. 72, pp. 101126, 2022, doi: 10.1016/j.stueduc.2022.101126.
[7] M. İlhan, N. Güler, G. T. Teker, and Ö. Ergenekon, "The effects of reverse items on psychometric properties and respondents' scale scores according to different item reversal strategies," *Int. J. Assess. Tools Educ*., vol. 11, no. 1, pp. 20–38, 2024, doi: 10.21449/ijate.1345549.
[8] C. Tang, B. Yang, and H. Tian, "Examination of the wording effect in the new ecological paradigm scale in China: a bi-factor modeling approach," *Curr. Psychol*., vol. 43, no. 7, pp. 5887–5900, 2024, doi: 10.1007/s12144-023-04801-z.
[9] J. García-Fernández, Á. Postigo, M. Cuesta, C. González-Nuevo, Á. Menéndez-Aller, and E. García-Cueto, "To be Direct or not: Reversing likert response format items," *Span. J. Psychol*., vol. 25, p. e24, Oct. 2022, doi: 10.1017/SJP.2022.20.
[10] F. A. Setiawati, S. R. Nurhayati, R. N. Amelia, and A. A. Darojat, "Study on the threats of reverse-worded items to the psychometric properties of the marital quality scale, *The Open Psychology Journal*, vol. 15, no. 1, pp. 1–8, 2022, doi: 10.2174/18743501-v15-e2208150.
[11] C. C. Koutsogiorgi and M. P. Michaelides, "Response tendencies due to item wording using eye-tracking methodology accounting for individual differences and item characteristics," *Behav. Res. Methods*, vol. 54, no. 5, pp. 2252–2270, 2022, doi: 10.3758/s13428-021-01719-x.
[12] D. Elek, H. Cígler, D. J. Grüning, and S. Ježek, "Advancing the psychometrics of reverse-keyed items: enriching cognitive theory by a logical and linguistic perspective," *Front. Psychol*., vol. 16, 2025, doi: 10.3389/fpsyg.2025.1684612.
[13] F. Antoniou and M. H. Alghamdi, "Confidence in mathematics is confounded by responses to reverse-coded items," *Front. Psychol*., vol. 15, 2024, doi: 10.3389/fpsyg.2024.1489054.
[14] S. Chen and B. Wei, "Development and validation of an instrument to measure high school students' science identity in science learning," *Research in Science Education*, vol. 52, no. 11, pp. 111-126, 2020, doi: 10.1007/s11165-020-09932-y.
[15] L. Avraamidou, "Science identity as a landscape of becoming: rethinking recognition and emotions through an intersectionality lens," *Cult. Stud. Sci. Educ*., vol. 15, no. 2, pp. 323–345, 2020, doi: 10.1007/s11422-019-09954-7.
[16] A. Venta et al., "Reverse-Coded items do not work in Spanish: Data from four samples using established measures," *Front. Psychol*., vol. 13, 2022, doi: 10.3389/fpsyg.2022.828037.
[17] B. Zeng, M. Jeon, and H. Wen, "How does item wording affect participants' responses in Likert scale? Evidence from IRT analysis," *Front. Psychol*., vol. 15, 2024, doi: 10.3389/fpsyg.2024.1304870.
[18] R. Komperda, K. N. Hosbein, and J. Barbera, "Evaluation of the influence of wording changes and course type on motivation instrument functioning in chemistry," *Chem. Educ. Res. Pract*., vol. 19, no. 1, pp. 184–198, 2017, doi: 10.1039/C7RP00181A.
[19] A. Rodriguez, S. P. Reise, and M. G. Haviland, "Evaluating bifactor models: Calculating and interpreting statistical indices.," Psychol. Methods, vol. 21, no. 2, pp. 137–150, 2016, doi: 10.1037/met0000045.
[20] M. Prokofieva, D. Zarate, A. Parker, O. Palikara, and V. Stavropoulos, "Exploratory structural equation modeling: a streamlined step by step approach using the R Project software," *BMC Psychiatry*, vol. 23, no. 1, p. 546, 2023, doi: 10.1186/s12888-023-05028-9.
[21] V. Swami, C. Maïano, and A. J. S. Morin, "A guide to exploratory structural equation modeling (ESEM) and bifactor-ESEM in body image research," *Body Image*, vol. 47, pp. 101641, 2023, doi: 10.1016/j.bodyim.2023.101641.
[22] J. Koran, "Indicators per factor in confirmatory factor analysis: more is not always better," *Struct. Equ. Model. A Multidiscip. J*., vol. 27, no. 5, pp. 765–772, 2020, doi: 10.1080/10705511.2019.1706527.
[23] T. A. Kyriazos, "Applied psychometrics: Sample size and sample power considerations in factor analysis (EFA, CFA) and SEM in general," *Psychology*, vol. 09, no. 08, pp. 2207–2230, 2018, doi: 10.4236/psych.2018.98126.
[24] S. Liu, S. Xu, Q. Li, H. Xiao, and S. Zhou, "Development and validation of an instrument to assess students ' science , technology , engineering , and mathematics identity," *Phys. Rev. Phys. Educ. Res*., vol. 19, no. 1, p. 10138, 2023, doi: 10.1103/PhysRevPhysEducRes.19.010138.
[25] J. Suárez-Álvarez, I. Pedrosa, L. Lozano, E. García-Cueto, M. Cuesta, and J. Muñiz, "Using reversed items in Likert scales: A questionable practice," *Psicothema*, vol. 2, no. 30, pp. 149–158, May 2018, doi: 10.7334/psicothema2018.33.
[26] N. Menold, "How Do Reverse-keyed Items in Inventories Affect Measurement Quality and Information Processing?," *Field methods*, vol. 32, no. 2, pp. 140–158, May 2020, doi: 10.1177/1525822X19890827.
[27] F. Kiwanuka, J. Kopra, N. Sak-Dankosky, R. C. Nanyonga, and T. Kvist, "Polychoric Correlation With Ordinal Data in Nursing Research," *Nurs. Res*., vol. 71, no. 6, pp. 469–476, Nov. 2022, doi: 10.1097/NNR.0000000000000614.
[28] S. Lim and S. Jahng, "Determining the number of factors using parallel analysis and its recent variants.," *Psychol. Methods*, vol. 24, no. 4, pp. 452–467, 2019, doi: 10.1037/met0000230.
[29] C. J. Gaskin and B. Happell, "On exploratory factor analysis: A review of recent evidence, an assessment of current

practice, and recommendations for future use," *Int. J. Nurs. Stud*., vol. 51, no. 3, pp. 511–521, 2014, doi: 10.1016/j.ijnurstu.2013.10.005.
[30] J. W. Osborne, "What is rotating in exploratory factor analysis?," *Pract. Assessment, Res. Eval*., vol. 20, no. 2, pp. 1–7, 2015, doi: 10.7275/hb2g-m060.
[31] P. Rogers, "Best practices for your confirmatory factor analysis: A JASP and lavaan tutorial," *Behav. Res. Methods*, vol. 56, no. 7, pp. 6634–6654, 2024, doi: 10.3758/s13428-024-02375-7.
[32] D. Shi, C. DiStefano, A. Maydeu-Olivares, and T. Lee, "Evaluating SEM model fit with small degrees of freedom," *Multivariate Behav. Res*., vol. 57, no. 2–3, pp. 179–207, 2022, doi: 10.1080/00273171.2020.1868965.
[33] D. Shi and A. Maydeu-Olivares, "The effect of estimation methods on SEM fit indices," *Educ. Psychol. Meas*., vol. 80, no. 3, pp. 421–445, 2020, doi: 10.1177/0013164419885164.
[34] S. P. Reise, W. Bonifay, and M. G. Haviland, "Bifactor modelling and the evaluation of scale scores," *The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development*, pp. 675–707, 2018, doi: 10.1002/9781118489772.ch22.
[35] K. S. Taber, "The use of cronbach's alpha when developing and reporting research instruments in science education," *Res. Sci. Educ*., vol. 48, no. 6, pp. 1273–1296, 2018, doi: 10.1007/s11165-016-9602-2.
[36] J. Wang, X. Xin, Y. Huo, Y. Li, Y. Han, and F. Kong, "Bifactor modelling, reliability, and validity of the material values scale in Chinese youth," *Psychol. Rep*., vol. 127, no. 1, pp. 465–484, 2024, doi: 10.1177/00332941221114407.
[37] M. S. Bartlett, "A note on the multiplying factors for various χ2 approximations," *J. R. Stat. Soc. Ser. B Stat. Methodol*., vol. 16, no. 2, pp. 296–298, 1954, doi: 10.1111/j.2517-6161.1954.tb00174.x.
[38] H. F. Kaiser, "An index of factorial simplicity," *Psychometrika,* vol. 39, no. 1. 1974. doi: 10.1007/BF02291575.
[39] M. W. Watkins, "Exploratory factor analysis: A guide to best practice," *J. Black Psychol*., vol. 44, no. 3, pp. 219–246, 2018, doi: 10.1177/0095798418771807.
[40] T. Zhang, C. Yin, Y. Geng, Y. Zhou, S. Sun, and F. Tang, "Development and validation of psychological contract scale for hospital pharmacists," *J. Multidiscip. Healthc*., vol. 13, pp. 1433–1442, 2020, doi: 10.2147/JMDH.S270030.
[41] C.-H. Li, "Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares," *Behav. Res. Methods*, vol. 48, no. 3, pp. 936–949, 2016, doi: 10.3758/s13428-015-0619-7.
[42] J. Revuelta, C. Ximénez, and N. Minaya, "Overfactoring in rating scale data: A comparison between factor analysis and item response theory," *Front. Psychol*., vol. 13, 2022, doi: 10.3389/fpsyg.2022.982137.
[43] P. J. Ferrando and U. Lorenzo-Seva, "Assessing the quality and appropriateness of factor solutions and factor score estimates in exploratory item factor analysis," *Educ. Psychol. Meas.*, vol. 78, no. 5, pp. 762–780, 2018, doi: 10.1177/0013164417719308.
[44] W. R. da Silva, G. S. Donofre, A. N. Neves, J. Marôco, P. A. Teixeira, and J. A. D. B. Campos, "Investigating method effects associated with the wording direction of items of the social physique anxiety scale," *Eat. Weight Disord. - Stud. Anorexia, Bulim. Obes*., vol. 27, no. 7, pp. 2857–2867, 2022, doi: 10.1007/s40519-022-01439-x.
[45] S. Savahl, F. Casas, and S. Adams, "Considering a bifactor model of children's subjective well-being using a multinational sample," *Child Indic. Res*., vol. 16, no. 6, pp. 2253–2278, 2023, doi: 10.1007/s12187-023-10058-6.
[46] C. C. S. Kam, "Why do regular and reversed items load on separate factors? response difficulty vs. item extremity," *Educ. Psychol. Meas*., vol. 83, no. 6, pp. 1085–1112, 2023, doi: 10.1177/00131644221143972.
[47] M. Fokkema and S. Greiff, "How performing PCA and CFA on the same data equals trouble," Eur. *J. Psychol. Assess*., vol. 33, no. 6, pp. 399–402, Nov. 2017, doi: 10.1027/1015-5759/a000460.
[48] I. Etikan, "Comparison of convenience sampling and purposive sampling," *Am. J. Theor. Appl. Stat*., vol. 5, no. 1, p. 1, 2016, doi: 10.11648/j.ajtas.20160501.11.
[49] G. D. Valenti, R. Bottaro, and P. Faraci, "Assessing the two sources of construct-relevant psychometric multidimensionality of the nomophobia questionnaire: The integrated framework of bifactor exploratory structural equation modeling," *Eval. Health Prof*., vol. 47, no. 1, pp. 52–65, 2024, doi: 10.1177/01632787231203380.
[50] R. E. Davis, S. Lee, T. P. Johnson, W. Yu, L. I. Reyes, and J. F. Thrasher, "Individual-level cultural factors and use of survey response styles among latino survey respondents," *Hisp. J. Behav. Sci.,* vol. 44, no. 3, pp. 216–242, 2023, doi: 10.1177/07399863231183023.
[51] A. Alamer, "Exploratory structural equation modeling (ESEM) and bifactor ESEM for construct validation purposes: Guidelines and applied example," *Res. Methods Appl. Linguist*., vol. 1, no. 1, pp. 100005, 2022, doi: 10.1016/j.rmal.2022.100005.
[52] D. Bolt, Y. C. Wang, R. H. Meyer, and L. Pier, "An IRT mixture model for rating scale confusion associated with negatively worded items in measures of social-emotional learning," *Appl. Meas. Educ*., vol. 33, no. 4, pp. 331–348, 2020, doi: 10.1080/08957347.2020.1789140.