JOURNAL OF CHEMICAL EDUCATION

Examining the Psychometric Properties of the Chemistry Self-Concept Inventory Using Rasch Modeling

Stephanie M. Werner, Ying Chen, and Mike Stieff*

Cite This: J. Chem. Educ. 2021, 98, 3412–3420

Read Online

ACCESS | Metrics & More | Article Recommendations

Downloaded via UNIV OF ILLINOIS CHICAGO on December 17, 2024 at 21:23:38 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

ABSTRACT: The Chemistry Self-Concept Inventory (CSCI) is a widely used instrument within chemistry education research. Yet, agreement on its overall reliability and validity is lacking, and psychometric analyses of the instrument remain outstanding. This study examined the psychometric properties of the subscale and item function of the CSCI on 1140 high school chemistry students in the midwestern United States. Using exploratory factor analysis, we found that a five-factor model, distinct from previous studies' models, fit our data best. Using this model in Rasch analysis, we found several items within the Mathematics subscale that showed unusual item difficulty, significant misfit, and low item discrimination. Implications for future use of the CSCI are discussed.

KEYWORDS: *High School/Introductory Chemistry, First-Year Undergraduate/General, Chemical Education Research, Testing/Assessment*

FEATURE: Chemical Education Research

INTRODUCTION

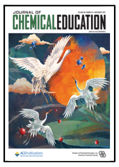
Affective factors that hinder or facilitate learning in the chemistry classroom are of increasing interest to the chemistry education research community.[1] Among these, students' self-concept with respect to chemistry learning has been identified as a contributing factor to persistence and achievement in the discipline. Whereas self-efficacy refers to perception of their own skills in a prospective manner, self-concept is defined as an individual's perception of themselves as a learner resultant from self-evaluation of their competencies in a domain.[2,3] Chemistry self-concept, specifically, has been studied along with other attitudinal measures as a predictor of student persistence and achievement in chemistry. For example, students' with a high chemistry self-concept are more likely to report autonomous learning strategies[4] and achieve higher retention rates.[5] More importantly, interventions that employ active learning techniques in college chemistry demonstrate that chemistry self-concept is sensitive to instructional conditions and can improve over time.[4−7] Such research demonstrates an important role for affective factors in predicting student success in chemistry and provides a target for improvement beyond the cognitive factors most commonly studied by the chemistry education research community.

Given the potential role of self-concept for promoting learning and persistence in chemistry, increased effort to provide valid interpretations of research instruments capable of measuring changes in chemistry self-concept is needed.[5,8] Chemistry self-concept is routinely measured by the Chemistry Self-Concept Inventory (CSCI). The primary context for validation studies of the CSCI has included convenience samples drawn from college student populations. While the measure appears to perform similarly among the sampled populations according to reported reliability statistics and factor analyses,[6,9] the validity of employing the measure in noncollege populations is uncertain. This makes the use of the CSCI in noncollege populations suspect as the psychometric properties of attitudinal measures are known to vary among populations with different demographics and appear particularly sensitive to social and cultural differences.[10]

Earlier research using the CSCI has demonstrated such sensitivity. In one study using the CSCI with students sampled from secondary chemistry classrooms in a rural setting, the CSCI appeared to perform distinctly from prior administrations with college students.[11] The five subscales identified in the original validation studies of the instrument were not evident among the sampled population. In fact, exploratory factor analysis identified only one subscale identical to the original CSCI validation study. Moreover, six constructs were proposed as opposed to the original five. Although the variable performance of the instrument between the two populations is unsurprising given other examples of such variability,[10] an explanation for the differing performance of the CSCI is lacking. While differences in chemistry self-concept on the instrument are likely due to maturation and history effects, the lack of consistency in the underlying constructs suggests that alternative analytical methods are necessary to explore the degree of validity and reliability in studies using the CSCI.

Here, we explore the psychometric properties of the CSCI by combining EFA with Rasch modeling. Although it is a

Received: April 14, 2021
Revised: September 17, 2021
Published: October 13, 2021



© 2021 American Chemical Society and Division of Chemical Education, Inc.

https://doi.org/10.1021/acs.jchemed.1c00436

powerful technique, exploratory factor analysis is highly subjective and relies on the researcher for interpretation and theoretical implications for decision making. Rasch modeling provides a complementary approach that can improve researcher interpretations of a measure's performance. Importantly, EFA can help to determine the factor structure according to participant responses and point to underlying constructs for further examination via Rasch modeling. Once these constructs are established, Rasch modeling permits an item-level analysis of the instrument to evaluate the scores' reliability and interpretation.[12,13] By combining these approaches, we examined whether the content of the instrument represents the magnitude of the construct measured in the CSCI. That is to say, the Rasch model provides a statistical framework against which researchers can compare empirical data to assess the CSCI's capacity to quantify and rank students' levels of self-concept. It also gives valuable information about construct theory to motivate additional studies into which items might need modification or removal. Such a fine-grained analysis of the CSCI has not been previously conducted to inform the research community about the discriminability and convergence of the CSCI items for analyzing chemistry self-concept.

■ RESEARCH QUESTIONS

The research questions for this study are as follows:

1. How is the reliability and validity of the CSCI as an instrument for measuring chemistry self-concept among high school chemistry students?
2. How do the CSCI items function for high school chemistry students based on Rasch analysis?

■ METHODS

This study was conducted with the approval of the Institutional Review Board of the University of Illinois—Chicago.

Sample

This data set included students from a larger research project exploring the efficacy of a high school chemistry curriculum. Students were removed from the data set if they did not complete all survey items (13%). The data set included complete survey responses from 1140 high school chemistry students that were all in or near a large Midwestern urban center.

Instrument

The CSCI is a 40-item survey measuring self-concept in several domains.[9] A student's self-concept is their belief about their abilities in general or specific domains of knowledge. The CSCI was adapted to a 5-point survey, with students rating items from "very inaccurate" to "very accurate". Students within this data set took the CSCI at the end of the academic year as part of a post-test measure. Using data from post-test administration is consistent with previous administrations and the development of self-concept, which states that self-concept is not fully formed until partway through the academic year.[9,11,14,15]

Data Analyses

The main purpose of this study was to examine the appropriateness of the items and internal structure of the constructs that the CSCI measures. In order to answer the research questions, an exploratory factor analysis was first conducted to evaluate the overall factor structure of the CSCI. Next, a Rasch analysis tested the construct validity of each item from the subscales defined by the EFA.

Exploratory Factor Analysis. EFA is a statistical method that allows us to ensure the reliability of the instrument by identifying inappropriate items within subscales and providing statistical evidence to remove inappropriate items. Further, this analysis also identifies the dimensionality of the constructs measured in the instrument through examining the relationships between factors and items.[16] In this study, we used parallel analysis (PA) to decide how many factors to retain when applying EFA. PA is a method based on Monte Carlo (MC) simulation techniques that compare the observed eigenvalues in simulated data with actual data.[17] Psychometricians in education endorse this approach for factor determination as PA is an accurate and robust method with less variability and sensitivity to a number of factors.[18−20]

The analysis of the CSCI for our high school sample used the *psych* package[21] in RStudio version 1.41717.[22]

Rasch Analysis. After the factor analysis, this study continued with Rasch modeling to conduct an item-level analysis on the CSCI. Generally speaking, Rasch models estimate the latent traits of participants and item characteristics based on the raw score. One benefit of Rasch modeling is that both person and item parameters are placed on a single standardized scale (logit scale) for direct comparison. The model assumes that the estimation of both parameters is invariant to the sample used, which allows unbiased and stable estimation of the item parameters.[23]

The Rasch Rating Scale model used in the current study is a derivation from the original dichotomous Rasch model.[24] It has been widely used to evaluate and investigate psychometric properties of polytomous scales such as Likert-type scales.[25] The model denoting the probability of person n with self-concept θ that endorses rating scale category k on a particular item i as opposed to endorsing rating scale $k - 1$ (i.e., the person chose rating scale category k "agree" over rating scale category $k - 1$ "neutral") is explained by this linear additive equation:

$$\ln\left[\frac{P_{n_i(x=k)}}{P_{n_i(x=k-1)}}\right] = \theta_n - \delta_i - \tau_k \tag{1}$$

Here, θ is the student's self-concept estimation, δ is the difficulty estimation of item i, and τ are the thresholds which are estimated empirically for the whole set of instruments. The student self-concept estimation refers to the student's tendency to endorse or agree with the CSCI overall, whereas the item difficulty estimation refers to the difficulty level for students to endorse a particular item. The rating scale model assumes that all threshold structure (distances among all rating scales) is constant across all items. Parameters for this model were estimated under joint maximum likelihood estimation (JMLE) in WINSTEPS 3.68.2.[26] The model fit statistics and item information were used to identify item performance and subscale information.

■ RESULTS

Exploratory Factor Analysis

Before preceding to execute exploratory factor analysis, data was screened for multivariate assumptions (e.g., normality, linearity, homogeneity and homoscedasticity), and the analysis suggested that all assumptions were met. Further, Bartlett's

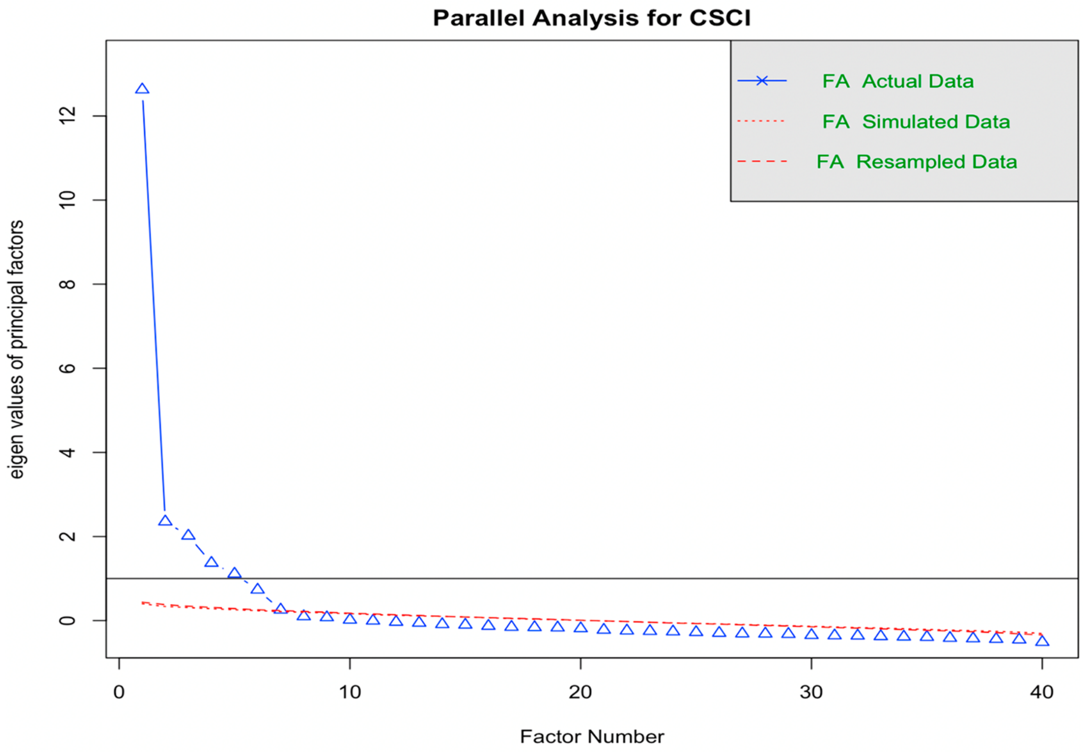


Figure 1. Scree plot shows the eigenvalue distribution obtained from actual CSCI data (blue line with triangles) vs randomly generated eigenvalue distribution from the parallel analysis (PA) simulation (red dashed lines). Each point on the blue line that lies above the simulated data line is a factor to extract. Thus, this plot shows an estimate of the number of factors to extract by looking at the number of points before the actual data and simulated data lines overlap. In this case, there are seven factors (the seventh triangle is slightly above the red lines). Note: The factor analysis (FA) simulated data and resampled data are overlapping in this plot, which indicates that both sampling algorithms yield comparable data for PA.

Test of Sphericity was conducted to determine if the correlations among all items were sufficiently large for the EFA; the result indicated correlation adequacy (X2(780) = 20916.43, $p < 0.001$). Next, the Kaiser−Meyer−Olkin (KMO) test was used to verify the sampling appropriate for the analysis. The KMO test indicated good sampling adequacy (MSA = 0.95), which aims for a higher value close to 1.[27−29] Given these overall indicators, a factor analysis was deemed to be suitable with CSCI items.

The parallel analysis (Figure 1) simulated eigenvalues from the raw data set along with the mean eigenvalue from the Monte Carlo simulations and random sampling simulation. However, in our case, the last factor (seventh) lies very close to the simulated factor analysis (FA) and resampled data lines. Thus, additional analyses were needed to further assess the possibility of a seven-factor solution for interpretability.

A more conservative retention criterion is to use the 95th percentile of eigenvalues generated from the MC simulated data,[18,30] which suggested seven factors (the vertical dashed line in Figure 2) for our data. According to the result from the PA, a factor analysis with maximum likelihood estimation was then executed. The analysis was performed with direct oblimin rotation because of the expected factor correlation. This model achieved simple structures with most items loading onto one factor.

The factor loadings for a seven-factor solution are shown in Table 1. After testing all 40 questions, six items (1, 11, 12, 21, 27, 38) split across several factors using the criterion that loadings must be greater than the absolute value of 0.30.[31] Items split across more than one factor were flagged for further analyses against the CSCI conceptual framework. Of the six items that cross-loaded, three items (1, 21, and 27) were placed in the factor with which they had a higher loading. For example, item 1 had a loading of 0.384 in factor 2 and a loading of 0.351 in factor 6. As for the remaining cross-loaded items (11, 12, and 38), these items were placed into a factor that had a lower factor loading compared to the other factor. We decided to move item 38 to factor 3 as opposed to 7 for several reasons. First, if this item remained in factor 7, it would be the only item in the factor, which is highly controversial in test design as a single item (or even pairs) is insufficient to meaningfully measure a relational construct.[32−34] Second, we considered the theoretical implications for placing an item into a particular factor loading.[32] For instance, we thought item 38 (*I could never achieve academic honors, even if I worked harder*) measured more of a student's belief in their capability and better aligned with similar capability items (as seen in factor 3 where items asked students about their general academic ability). Similar theoretical implications were considered for items 11 and 12. Item 11 (*I wish I had more imagination and originality*) was placed with other items that had item stems measuring creativity, ideas, imagination, and curiosity. Item 12 (*I find chemistry concepts interesting and challenging*) was placed with all other chemistry self-concept items as opposed to other items measuring general academic interest. Once all cross-loaded items were placed into a single factor for practical and theoretical reasons, factors 6 and 7 no longer contained any items, resulting in a reduced solution with five factors (i.e., factors 1−5). Further, items 3 (*I am never able to think up answers to problems that have not already been figured out*) and 35 (*I would have no interest in being an inventor*) did not load on to any factor.

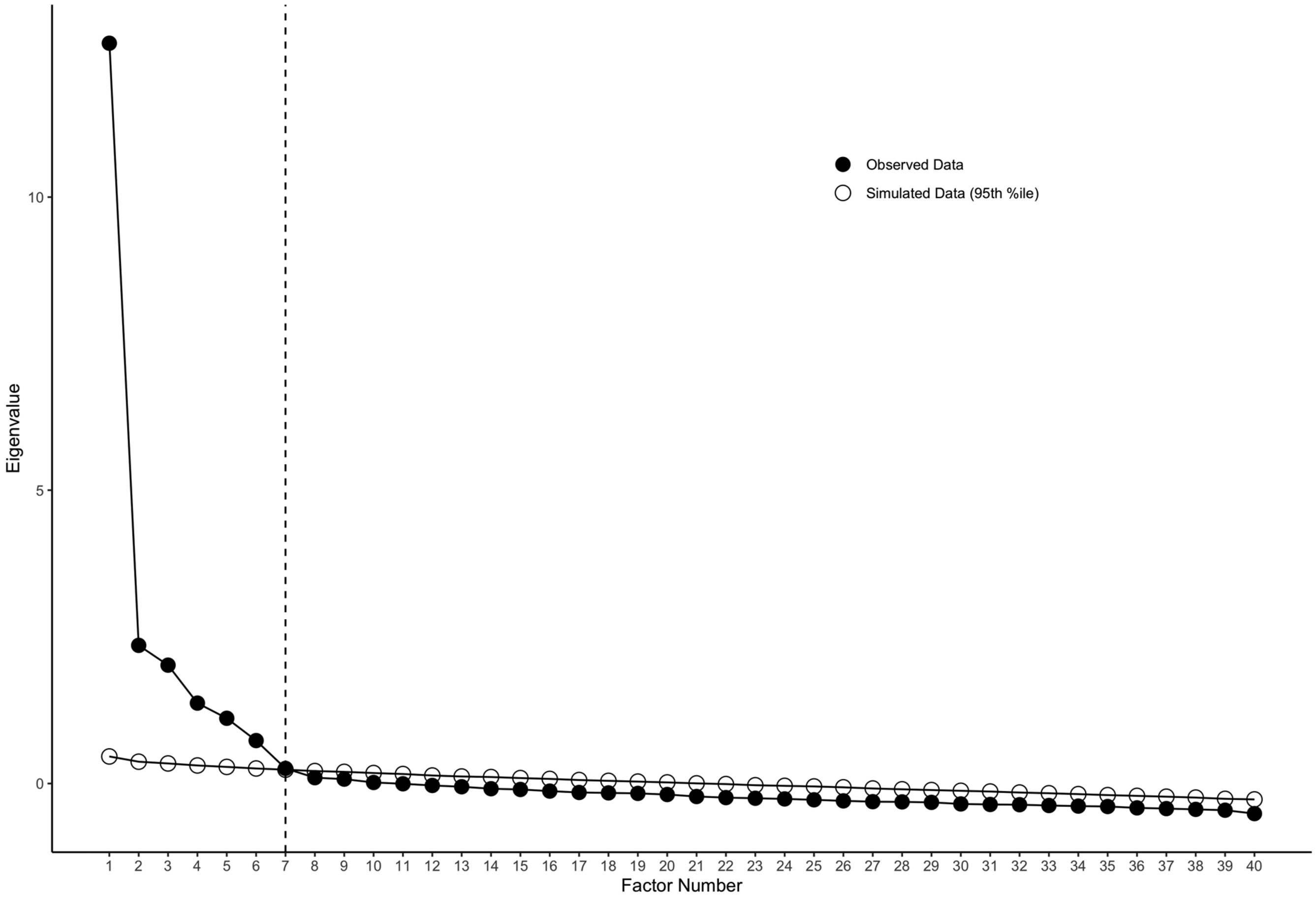


Figure 2. Decision plot shows the actual (observed) eigenvalues drawn from the CSCI data, as well as the 95th percentile eigenvalues drawn from the parallel analysis simulated data. The vertical dashed line indicates the factor retention with 95th percentile criterion by counting the number of factors on and before the dashed line. Here, the dashed line crosses at seven factors, so this more conservative approach of the 95th percentile criterion suggests a solution of seven factors.

The resulting five factors were labeled as Chemistry self-concept, Mathematics self-concept, Academic Capability self-concept, Academic Enjoyment self-concept, and Creativity self-concept. Cronbach's α values (Table 2) showed reasonably good internal consistency reliabilities of the factors, or subscales, extracted in the EFA.[35,36] However, the Creativity subscale showed low reliability with a value of 0.32. Further investigation on this subscale found that items 11 and 27 had negative item-total correlations. After removing these items from the Creativity self-concept scale, there was a marked improvement in Cronbach's α value (0.68) when considering a subscale with five items. Thus, removing these two items was justified.

Rasch Analysis

Our second objective for this study was to investigate the psychometric properties of the CSCI items and subscales using Rasch modeling. To achieve this goal, we conducted a set of item performance analyses using consecutive modeling for each subscale. The major justifications for using a consecutive model are as follows: (1) the students' self-concept measured by the CSCI is multidimensional, and the assumption of unidimensionality does not hold according to our previous factor analytic procedures as well as previous studies;[9,11] (2) application of composited Rasch model on a multidimensional instrument can bias parameter estimation and result in significant test information loss on the subscale level;[37] (3) a multidimensional IRT model is technically and computationally complex;[38] consequently, the output is difficult to interpret as seeking unidimensional instrument is not the purpose of this study.

In the consecutive approach, the scores on items associated with the five subscales of the CSCI are treated as five different indicators and are modeled separately into individual analyses. In other words, the consecutive approach is simply a unidimensional model repeated five times using a subset of items included on a given subscale. Because each analysis only includes items defined by each latent domain, the effect of relationships between the five subscales is removed. Thus, this approach has the advantage of yielding self-determination estimates and smaller standard error of measurement for all five subscales. Compared with the composite approach, the consecutive approach retains more item-level information for each subscale.[39]

Table 3 includes the item difficulty, fit statistics, standard error, point measurement correlation, and estimated discrimination for each item of the CSCI using the five-subscale model after removing item 11 and item 27. Each item statistic is discussed in detail below.

Item Difficulty. We examined the item difficulty measures and the distribution to warrant the overall performance of the CSCI. Higher difficulty estimates (greater values) indicate that a particular item is less likely to be endorsed by students (more likely to choose a "very inaccurate" rating) and vice versa. The

Table 1. Factor Loadings for CSCI[a]

Item Number	Item	Factor 1	2	3	4	5	6	7
	Chemistry Self-Concept							
4*	I have never been excited about chemistry.	0.442						
8	I participate confidently in discussions with school friends about chemical topics.	0.470						
12	I find chemistry concepts interesting and challenging.	0.340					0.476	
16	When I run into chemical topics in my courses, I always do well on that part.	0.712						
20*	I would hesitate to enroll in courses that involve chemistry.	0.585						
24	I am quite good at dealing with chemical ideas.	0.715						
28*	Chemistry intimidates me.	0.603						
32*	I have always had difficulty understanding arguments that require chemical knowledge.	0.557						
36	I have always done better in courses that involve chemistry than in most courses.	0.766						
40*	I have trouble understanding anything based on chemistry.	0.606						
	Mathematics Self-Concept							
1	I find many math problems interesting and challenging.		0.384				0.351	
5*	I have hesitated to take courses that involve math.		0.590					
9	I have generally done better in math courses than in other courses.		0.876					
13*	Math makes me feel inadequate.		0.599					
17	I am quite good at math.		0.852					
19*	I'm not much good at problem solving.		0.313					
21*	I have trouble understanding anything based on math.		0.641					0.329
25	I have always done well in math classes.		0.782					
29*	I never do well on tests that require math reasoning.		0.546					
33	At school, my friends always come to me for help in math.		0.552					
37*	I have never been very excited about math.		0.683					
	Academic Capability Self-Concept							
14*	I have trouble with most academic subjects.			0.344				
18	I'm good at most academic subjects.			0.514				
26	I learn quickly in most academic subjects.			0.319				
34	I get good marks in most academic subjects.			0.550				
38*	I could never achieve academic honors, even if I worked harder.			0.319				0.329
	Academic Enjoyment Self-Concept							
2	I enjoy doing work for most academic subjects.				0.617			
6*	I hate studying many academic subjects.				0.722			
10	I like most academic subjects.				0.623			
22*	I'm not particularly interested in most academic subjects.				0.775			
30*	I hate most academic subjects.				0.757			
	Creativity Self-Concept							
7	I am good at combining ideas in ways that others have not tried.					0.629		
11*	I wish I had more imagination and originality.					−0.334	0.483	
15	I enjoy working out new ways of solving problems.					0.323		
23	I have a lot of intellectual curiosity.					0.488		
27*	I am not very original in my ideas, thoughts, and actions.					−0.463	0.367	
31	I am an imaginative person.					0.594		
39	I can often see better ways of doing routine tasks.					0.496		

[a] Factor loadings have been sorted, and loadings that fell below the threshold (|0.3|) in Table 1 were suppressed. * Indicates reverse coded items.

Table 2. Cronbach's α Values

Subscale	Cronbach's α	Number of Items
Mathematics self-concept	0.91	10
Chemistry self-concept	0.88	11
Academic Capability self-concept	0.81	5
Academic Enjoyment self-concept	0.84	5
Creativity self-concept	0.32	7
Creativity self-concept (without items 11 and 27)	0.68	5

relatively most difficult item was item 36 (*I have always done better in courses that involve chemistry than in most courses*; 0.98), which means students were more likely to disagree with this item. The relatively easiest item (student *most* endorsed) was item 21 (*I have trouble understanding anything based on math*; −0.58). Overall, no items in any subscale clustered at extreme low or high item difficulty values, where the item difficulty spectrum is within ±1 range. Thus, we did not observe significant flooring or ceiling effects in this analysis. Notably, reverse-coded items (marked with asterisk) were easier for students to endorse, apart from item 6 (*I hate studying many*

Table 3. Item Fit Statistics from Rasch Analysis

		MNSQ				
Item	Difficulty	Infit	Outfit	Standard Error	PT-Measure Corr	Estim Discr
Chemistry Self-Concept						
4*	−0.49	1.07	1.07	0.03	0.65	0.93
8	0.41	1.21	1.29	0.04	0.61	0.69
12	−0.39	1.24	1.26	0.04	0.59	0.72
16	0.37	0.84	0.83	0.04	0.72	1.18
20*	−0.08	0.96	0.94	0.03	0.7	1.07
24	0.26	0.81	0.81	0.04	0.74	1.23
28*	−0.37	1.04	1.09	0.04	0.65	0.94
32*	−0.28	0.99	1	0.04	0.67	1.02
36	0.98	1.01	1.02	0.04	0.67	1
40*	−0.41	0.85	0.91	0.04	0.71	1.18
Mathematics Self-Concept						
1	0.02	1.5	1.55	0.04	0.55	0.41
5*	−0.14	1.07	1.08	0.03	0.69	0.94
9	0.34	0.83	0.84	0.03	0.76	1.22
13*	−0.09	1.02	1.06	0.04	0.69	0.97
17	0.08	0.62	0.65	0.04	0.8	1.46
19	−0.19	1.22	1.25	0.04	0.6	0.74
21*	−0.58	0.73	0.69	0.04	0.76	1.35
25	−0.01	0.71	0.69	0.04	0.78	1.38
29*	−0.35	0.97	0.98	0.04	0.7	1.05
33	0.82	1.35	1.46	0.04	0.62	0.57
37*	0.09	0.95	0.95	0.03	0.72	1.07
Academic Capability Self-Concept						
14*	−0.18	0.96	0.96	0.04	0.73	1.05
18	0.15	0.9	0.89	0.05	0.76	1.1
26	0.27	1.2	1.2	0.04	0.7	0.76
34	0.13	0.88	0.87	0.04	0.78	1.15
38*	−0.37	1.04	1.08	0.04	0.68	0.96
Academic Enjoyment Self-Concept						
2	0.61	1.17	1.17	0.04	0.74	0.81
6*	0.13	1.06	1.08	0.04	0.75	0.92
10	0.02	0.98	0.98	0.05	0.77	1.02
22*	−0.19	0.92	0.91	0.04	0.79	1.11
30*	−0.55	0.84	0.81	0.04	0.79	1.21
Creativity Self-Concept						
7	0.33	0.89	0.89	0.04	0.70	1.12
15	0.37	1.06	1.07	0.04	0.64	0.93
23	−0.38	0.87	0.86	0.04	0.68	1.15
31	−0.15	1.09	1.13	0.04	0.61	0.87
39	−0.16	1.05	1.04	0.04	0.62	0.94

academic subjects; 0.13). Items within four of the five subscales showed uniform measurement difficulty centered around zero. However, the Mathematics subscale showed item difficulty negatively skewed. After removing misfit items from this subscale, the remaining items were overall easier for students to endorse compared with the items from the other four subscales which were much harder for students to endorse.

Item Fit Statistics. In Rasch measurement, the mean square (MNSQ) fit measures are often used to determine whether individual items match the expected estimation of the Rasch model.[40] Infit and outfit statistics provide evidence about the construct validity of the instrument. An infit statistic is a weighted goodness-of-fit statistic, which is often used to diagnose unexpected patterns of responses on items (weighted more on unexpected response pattern when an item is close to the student's latent ability), whereas the outfit statistic is an unweighted goodness-of-fit statistic, signaling unexpected patterns from outliers (unexpected responses from students with extreme ability estimates).

However, with an expected value of 1.0, there is no rule-of-thumb of lower and upper limits for acceptable fit values. Lincare and Wright gave a detailed guideline on the cutoff points and suggested a range between 0.5 and 1.5 as acceptable.[41] Bond and Fox suggested that a narrower range of 0.6−1.4 should be used for the rating scale model.[42] In this study, considering that the CSCI was not used as a high stakes instrument but rather as a "run of the mill" instrument, we set the cutoff range to 0.7−1.3 as suggested by Wright and Linacre.[43] An item showing a fit value below the lower limits (i.e., overfit) suggests a nontrivial interdependence with another item, meaning predictable response patterns, whereas an item with a fit value above the upper limit (i.e., underfit) signals the item has more variance or noise than predicted by the Rasch model. Table 3 shows a good overall item fit across the subscales. However, item 1 (*I find many math problems interesting and challenging*; infit = 1.50; outfit = 1.55) and item 33 (*At school, my friends always come to me for help in math*; infit = 1.35; outfit = 1.46) showed an underfit, suggesting unusual response patterns. Conversely, item 17 (*I am quite good at math*; infit = 0.62; outfit = 0.65) exhibited more predictable patterns. Item 21 (*I have trouble understanding anything based on math*; infit = 0.73; outfit = 0.69) and item 25 (*I have always done well in math classes*; infit = 0.71; outfit = 0.69) showed similar issues but had a marginal overfit value. Additionally, residual correlation analysis highlighted potential local dependency of this pair of overfitted items (item 17 and 25, $r = 0.19$), even though standard reporting in the literature usually seeks residual correlation over the critical value of 0.2.[44]

Item Polarity. We calculated *Point of Measure Correlation* (*PT-Measure Corr*) to detect item polarity. It is intended to examine the conformity of the instrument and whether the construction of the constructs meets the measurement intentions.[45] If the value of *PT-Measure Corr* is above 0, this particular item measures the construct as it is intended. A high *PT-Measure Corr* value means that an item functions to distinguish between students' self-concepts. Similar to other correlation coefficient interpretations, the absolute value greater than 0.35 is generally considered a weak correlation, 0.36−0.67 a moderate correlation, and 0.68 or above a strong correlation.[46] The calculated point-measure correlations indicate all items in each subscale are pointing to a uniform direction and signal adequate coherence with values above 0.5.[46]

Item Discrimination. We investigated item performance via the Estimated Discrimination index. Though Rasch models assert item discrimination to be equal at a value of 1.0, in reality the item discrimination varies; therefore, the estimated item discrimination is considered as a type of fit statistic. The Estimated Discrimination is a posthoc estimation of item discrimination. When the discrimination parameter for an item is greater than 1.0, it indicates that this item could more effectively discriminate between students with high and low self-concept than expected for an item of this difficulty.[47] The results suggest that most items exhibited adequate discrimination, except for item 1 (*I find many math problems interesting and challenging*; 0.41) and item 33 (*At school, my friends always come to me for help in math*; 0.57). These values indicate the two items did not discriminate students' math self-concept as expected in our sample. Overall, we did not observe any item overdiscriminating compared to other items as a set or an extremely discriminating item in any subscale.

Item Reliability. Table 4 shows the summary statistics of item separation and reliability for each subscale. Item separation and reliability indices are two reproducibility statistics that are supplementary to evaluating the adequacy of the Rasch rating scale model. The item separation index examines the structure of how well the items are placed on the latent structure (i.e., the spread of items on the measured latent constructs). Typically, a separation value greater than 2 is desired.[48] The separation indices ranged from 5.20 to 12.16 for each subscale, which reveals an adequate structure and spacing on samples sharing similar self-concepts. The item reliability measures how well this instrument could effectively differentiate persons on the measured latent construct. Values greater than 0.7 are desired,[23] and item reliability values for each subscale are well above this threshold of acceptance. Overall, the item reliability and separation indices signal an outstanding psychometric quality for the CSCI subscales based on our five-factor model.

Table 4. Fit Criteria for Consecutive Approach

Subscale	Separation	Reliability	MNSQ Infit	MNSQ Outfit
Chemistry	12.16	0.99	1.00 (0.14)	1.02 (0.15)
Mathematics	9.04	0.99	1.00 (0.26)	1.02 (0.29)
Academic Capability	5.20	0.96	1.00 (0.12)	1.00 (0.13)
Academic Enjoyment	8.62	0.99	0.99 (0.11)	0.99 (0.13)
Creativity	7.52	0.98	1.00(0.09)	1.00 (0.10)

DISCUSSION

The EFA of the CSCI yielded a five-factor solution for urban/suburban high school chemistry students that included the following subscales: Chemistry self-concept, Mathematics self-concept, Academic capability self-concept, Academic enjoyment self-concept, and Creativity self-concept. Our factor solution, while showing some similarities to previous analyses of the CSCI,[9,11] is distinct in the item makeup of each subscale.

Through EFA, we found two issues with the CSCI at the subscale level. The first issue is that two items (3 and 35) did not belong to any subscale. Their factor loadings were below the threshold to be associated with any of the factors. This is problematic for a few reasons. First, if the primary use of the CSCI is to analyze survey responses at the subscale-level (which is more common than on the item-level), then these items are useless in the sense that they would not be used as comparative measures across students. Also, item 3 in particular has been found "factor-less" in previous studies of the CSCI. Including this study, there have been three studies when the CSCI has undergone EFA using a range of diverse participants, and twice, item 3 has resulted in no factor.[9] As for item 35, our study is the first instance of this item having no factor. In previous studies, item 35 has belonged in subscales containing different items each of time.[9,11] Thus, this item is more unpredictable in its associations with other items.

It is important to note that we found several items (items 1, 11, 12, 21, 27, 38) that cross-loaded onto more than one factor. In fact, each of these item stems ask respondents to evaluate multiple aspects (e.g., item 1: *I find many math problems**interesting and challenging***) or include language that is extreme (e.g., item 21: *I have trouble understanding**anything** based on math*). This language may be the reason they were cross-loaded. While cross-loading items may be cause for concern as to which factor they are truly measuring, theoretical considerations were made in an attempt to retain as many items as possibly to run the item-level Rasch analysis in the next step. Researchers may want to review students' interpretation of the language in these item stems to prevent cross-loading in the future.

The second issue found from the EFA is the reliability, or lack thereof, of the Creativity subscale. This subscale was not reliable (low Cronbach's α value) with all associated items. However, upon reviewing the items and removing two (items 11 and 27; both of which are reverse-coded), the reliability improved for the subscale. By removing these two items, we were able to retain a reliable Creativity subscale with five items as opposed to seven.

We used the five subscales resulting from the EFA to measure the psychometric properties of the items by examining item difficulty, fit, discrimination, and reliability via Rasch modeling. In terms of item difficulty, the Chemistry subscale showed a broad range of difficulty. In other words, the items are well-targeted to persons with varying levels of chemistry self-concept. The same was true for those subscales with fewer items (Academic Capability, Academic Enjoyment, and Creativity). However, we discovered that the Mathematics self-concept subscale exhibited a narrow range of measurement that was negatively skewed. Several items in this subscale could only produce reliable and precise student estimates for those with low and medium self-concept. Therefore, this subscale was not capable of measuring students with an above average Mathematics self-concept.

Results from the item fit and item discrimination analyses revealed two key issues in the Mathematics subscale. First, both item 1 and item 33 were underfit, which suggests the responses of these two items had more unexpected variances than the Rasch model expected. Further, these items had relatively low item discriminations, meaning they could not distinguish high and low self-concept students and are not measuring the latent construct effectively. One possibility for these issues is the ambiguous language used in these two items. The term "challenging" in item 1 (*I find many math problems interesting and challenging*) could either be construed in a positive or negative sense and give way to a conflicting response. Item 33 (*At school, my friends always come to me for help in math*) is the only item on the CSCI that asks the respondent to make a determination about a peer's self-concept or how peers view the respondent's self-concept. This change in perspective from the rest of the items on the inventory coupled with the underfit and low discrimination show that this item is out of place. The ambiguous language in both items could cause students to misunderstand the items and lead to disturbances in their response patterns. The disturbance in item responses is likely to cause inaccurate measurement. These items should be reconceptualized (i.e., reworded or revised) and retested to see if fits improve with less ambiguous wording.

The second issue arises from another set of items from the Mathematics subscale: items 17 (*I am quite good at math*) and 25 (*I have always done well in math classes*). The results from the item fit suggest these items operate in parallel with similar fit statistics, with both showing overfit, relatively high point of measure correlations, and their residual correlation suggest mutual dependence. These results support the claim that these items are similarly worded and are redundant within the same subscale. The use of repeated items has two significant and undesirable outcomes. The inclusion of parallel items will narrow the spectrum of the measurement and result in a lower degree of validity of the subscale.[49] Moreover, this redundancy could cause false high subscale reliability because higher homogeneity produces unwanted common factors of the scale.[48] To resolve this issue, either item could be dropped or reconceptualized and tested again to see if item fit improves.

Beyond the concerns already addressed for items within the Mathematics subscale, items within the other four subscales performed well for this high school sample. All other items make contributions to their respective latent constructs, showed no misfit, exhibited a good range of discrimination, and had difficulty estimates within acceptable ranges. At the subscale level, separation and reliability results indicated high internal consistency, similar to what was found for the EFA. Overall, the Chemistry, Academic Capability, Academic Enjoyment, and Creativity items function well with high school students. Several items in the Mathematics subscale should be reconceptualized and retested to see if individual items and the subscale as a whole can be improved.

LIMITATIONS

There are some limitations to this study that should be noted. First, demographic information on students was unavailable, and thus, a differential item functioning (DIF) analysis was impossible for Rasch analysis. Without this additional analysis, it is uncertain if the instrument is fair across various demographic groups.[50] If possible, future work should include this data in their analyses of this instrument. Second, we were unable to retest the instrument implementing the changes discussed above (i.e., reconceptualized items). Future studies may revise the instrument as proposed, administer it on a new sample, and reexamine the validity and reliability of the shortened instrument.

IMPLICATIONS

Considering that the CSCI has been widely used in chemistry education research, we offer several implications from our findings that should be considered in future work. First, negatively stated (or reverse-coded) items should not be used, especially with high school students who may be more likely to be confused by such wording. This finding is consistent with the existing literature that shows mixing positive and negative item phrasing leads to respondent confusion that may cause inaccurate measurement.[51,52] Further, our findings showed an association between reverse-coded items and low item difficulty (easy to endorse), whereas positively stated items showed good variation and range with item difficulty.

As discussed above, two items (3 and 35) did not belong to any factor in our analysis. Item 3 has been previously found to perform problematically with other studies using different student samples (based on age and chemistry experience). Thus, we propose this item be removed from future analyses and administrations. As for item 35, it is unclear if this item is problematic across various student demographics (especially since this item has been included in factors previously) and thus highlights the need for future administrators to complete exploratory or confirmatory factor analyses depending on the similarity of their respondents with previous studies.

Lastly, though the Mathematics subscale exhibited outstanding internal consistency through the EFA, the Rasch analysis revealed that several items showed unusual item difficulty, significant overfit, and low item discrimination. Because of the issues surrounding the Mathematics subscale, future work should investigate students' understanding of each item through surveys or cognitive interviews to determine how these items could be modified to improve the subscale and overall instrument.

AUTHOR INFORMATION

Corresponding Author

Mike Stieff – *Department of Chemistry, University of Illinois at Chicago, Chicago, Illinois 60607-7101, United States; Learning Sciences Research Institute, University of Illinois at Chicago, Chicago, Illinois 60607-7101, United States;* orcid.org/0000-0002-1639-891X; Email: mstieff@uic.edu

Authors

Stephanie M. Werner – *Department of Chemistry, University of Illinois at Chicago, Chicago, Illinois 60607-7101, United States;* orcid.org/0000-0002-5819-7331

Ying Chen – *Department of Educational Psychology, University of Illinois at Chicago, Chicago, Illinois 60607-7101, United States*

Complete contact information is available at:
https://pubs.acs.org/10.1021/acs.jchemed.1c00436

Notes

The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education.

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The research reported here was supported, in whole or in part, by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A170074 to the University of Illinois—Chicago. The authors express their appreciation to the chemistry teachers and students who participated in this project. The authors would also like to gratefully acknowledge the assistance of Dr. Yue Yin for guidance on analyses and Dr. Alison Castro Superfine, Rachel Yim, Melanie Snow, and Sidney Spurgeon for their thoughtful feedback on this manuscript.

REFERENCES

(1) Flaherty, A. A. A Review of Affective Chemistry Education Research and Its Implications for Future Research. *Chem. Educ. Res. Pract.* **2020**, *21*, 698–713.

(2) Marsh, H. W.; Walker, R.; Debus, R. Subject-Specific Components of Academic Self-Concept and Self-Efficacy. *Contemp. Educ. Psychol.* **1991**, *16* (4), 331–345.

(3) Beane, J. A.; Lipka, R. P. *Self-Concept, Self-Esteem, and the Curriculum*; Teachers College Press: New York, 1986.

(4) Chan, J. Y. K.; Bauer, C. F. Learning and Studying Strategies Used by General Chemistry Students with Different Affective Characteristics. *Chem. Educ. Res. Pract.* **2016**, *17*, 675.

(5) Lewis, S. E.; Shaw, J. L.; Heitz, J. O.; Webster, G. H. Attitude Counts: Self-Concept and Success in General Chemistry. *J. Chem. Educ.* **2009**, *86* (6), 744–749.

(6) Vincent-Ruz, P.; Meyer, T.; Roe, S. G.; Schunn, C. D. Short-Term and Long-Term Effects of POGIL in a Large-Enrollment General Chemistry Course. *J. Chem. Educ.* **2020**, *97* (5), 1228−1238.
(7) Chan, J. Y. K.; Bauer, C. F. Identifying At-Risk Students in General Chemistry via Cluster Analysis of Affective Characteristics. *J. Chem. Educ.* **2014**, *91* (9), 1417−1425.
(8) Arjoon, J. A.; Xu, X.; Lewis, J. E. Understanding the State of the Art for Measurement in Chemistry Education Research: Examining the Psychometric Evidence. *J. Chem. Educ.* **2013**, *90* (5), 536−545.
(9) Bauer, C. F. Beyond "Student Attitudes": Chemistry Self-Concept Inventory for Assessment of the Affective Component of Student Learning. *J. Chem. Educ.* **2005**, *82* (12), 1864−1870.
(10) Cooper, S. R.; Gonthier, C.; Barch, D. M.; Braver, T. S. The Role of Psychometrics in Individual Differences Research in Cognition: A Case Study of the AX-CPT. *Front. Psychol.* **2017**, *8*. DOI: 10.3389/fpsyg.2017.01482.
(11) Nielsen, S. E.; Yezierski, E. Exploring the Structure and Function of the Chemistry Self-Concept Inventory with High School Chemistry Students. *J. Chem. Educ.* **2015**, *92* (11), 1782−1789.
(12) Mui Lim, S.; Rodger, S.; Brown, T. Using Rasch Analysis to Establish the Construct Validity of Rehabilitation Assessment Tools. *Int. J. Ther. Rehabil.* **2009**, *16* (5), 251−260.
(13) Baghaei, P. The Rasch Model as a Construct Validity Tool. *Rasch Meas. Trans.* **2008**, *22*, 1145−1146.
(14) Bong, M.; Skaalvik, E. M. *Academic Self-Concept and Self-Efficacy: How Different Are They Really?*; 2002; Vol. *18*.
(15) Nieswandt, M. Student Affect and Conceptual Understanding in Learning Chemistry. *J. Res. Sci. Teach.* **2007**, *44* (7), 908−937.
(16) Netemeyer, R. G.; Bearden, W. O.; Sharma, S. *Scaling Procedures: Issues and Applications*; SAGE Publications, 2003.
(17) Horn, J. L. A Rationale and Test for the Number of Factors in Factor Analysis. *Psychometrika* **1965**, *30*, 179−185.
(18) Glorfeld, L. W. An Improvement on Horn's Parallel Analysis Methodology for Selecting the Correct Number of Factors to Retain. *Educ. Psychol. Meas.* **1995**, *55*, 377−393.
(19) Thompson, B.; Daniel, L. G. Factor Analytic Evidence for the Construct Validity of Scores: A Historical Overview and Some Guidelines. *Educ. Psychol. Meas.* **1996**, *56* (2), 197−208.
(20) Zwick, W. R.; Velicer, W. F. Comparison of Five Rules for Determining the Number of Components to Retain. *Psychol. Bull.* **1986**, *99*, 432−442.
(21) Revelle, W. Psych: Procedures for Personality and Psychological Research. 2017.
(22) RStudio Team. *RStudio: Integrated Development for R*; Boston, 2020.
(23) Embretson, S. E.; Reise, S. P. *Item Response Theory for Psychologists (Multivariate Applications Series)*; Lawrence Erlbaum Associates Publishers, 2000.
(24) Andrich, D. A Rating Formulation for Ordered Response Categories. *Psychometrika* **1978**, *43* (4), 561−573.
(25) Andrich, D. *Rasch Models for Measurement*; SAGE Publications: Beverly Hills, 1988.
(26) Linacre, J. M. *Winsteps Rasch Measurement Computer Program*; Beaverton, OR, 2009.
(27) Kaiser, H. F. A Second Generation Little Jiffy. *Psychometrika* **1970**, *35* (4), 401−415.
(28) Kaiser, H. F. An Index of Factorial Simplicity. *Psychometrika* **1974**, *39* (1), 31−36.
(29) Kaiser, H. F.; Rice, J. Little Jiffy, Mark Iv. *Educ. Psychol. Meas.* **1974**, *34* (1), 111−117.
(30) Harshman, R. A.; Reddon, J. R. Determining the Number of Factors by Comparing Real with Random Data: A Serious Flaw and Some Possible Corrections. *Classification Society of North America at Philadelphia* **1983**, 14−15.
(31) Howard, M. C. A Review of Exploratory Factor Analysis Decisions and Overview of Current Practices: What We Are Doing and How Can We Improve? *Int. J. Hum. Comput. Interact.* **2016**, *32* (1), 51−62.
(32) Ferguson, E.; Cox, T. Exploratory Factor Analysis: A Users' Guide. *Int. J. Sel. Assess.* **1993**, *1* (2), 84−94.
(33) Norton, R. Measuring Marital Quality: A Critical Look at the Dependent Variable. *J. Marriage Fam.* **1983**, *45* (1), 141−151.
(34) Petrescu, M. Marketing Research Using Single-Item Indicators in Structural Equation Models. *J. Mark. Anal.* **2013**, *1*, 99−117.
(35) DeVellis, R. F. *Scale Development: Theory and Applications*, 3rd ed.; SAGE Publications: CA, 2012.
(36) Cronbach, L. J. *Essentials of Psychological Testing*; Harper & Row: New York, 1970.
(37) Folk, V. G.; Green, B. F. Adaptive Estimation When the Unidimensionality Assumption of IRT Is Violated. *Appl. Psychol. Meas.* **1989**, *13* (4), 373−390.
(38) Hardouin, J.-B.; Mesbah, M. Clustering Binary Variables in Subscales Using an Extended Rasch Model and Akaike Information Criterion. *Commun. Stat. - Theory Methods* **2004**, *33* (6), 1277−1294.
(39) Baghaei, P. The Application of Multidimensional Rasch Models in Large Scale Assessment and Validation: An Empirical Example. *Electron. J. Res. Educ. Psychol.* **2012**, *10* (1), 1696−2095.
(40) Smith, A. B.; Rush, R.; Fallowfield, L. J.; Velikova, G.; Sharpe, M. Rasch Fit Statistics and Sample Size Considerations for Polytomous Data. *BMC Med. Res. Methodol.* **2008**, *8* (1), 1−11.
(41) Linacre, J. M.; Wright, B. Reasonable mean-square fit values. https://www.rasch.org/rmt/rmt83b.htm.
(42) Bond, T. G.; Fox, C. M. *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*; Lawrence Erlbaum Associates Publishers, 2001.
(43) Wright, B.; Linacre, J. M. Reasonable Mean-Square Fit Values. *Rasch Meas. Trans.* **1994**, *8*, 370−371.
(44) Chen, W.-H.; Thissen, D. Local Dependence Indexes for Item Pairs Using Item Response Theory. *J. Educ. Behav. Stat.* **1997**, *22* (3), 265−289.
(45) Boone, W. J.; Staver, J. R. Point Measure Correlation. In *Advances in Rasch Analyses in the Human Sciences*; Springer: Cham, Switzerland, 2020; pp 25−38.
(46) Othman, N. B.; Salleh, S. M.; Hussein, H.; Wahid, H. B. A. Assessing Construct Validity and Reliability of Competitiveness Scale Using Rasch Model Approach. *2014 WEI International Academic Conference Proceedings* **2014**, 113−120.
(47) Masters, G. N. Item Discrimination: When More Is Worse. *J. Educ. Meas.* **1988**, *25* (1), 15−29.
(48) Boone, W. J.; Staver, J. R.; Yale, M. S. Person Reliability, Item Reliability, and More. In *Rasch Analysis in the Human Sciences*; Springer: Dordrecht, 2014; pp 217−234. DOI: 10.1007/978-94-007-6857-4_10
(49) Bond, T. G.; Fox, C. M.; Lacey, H. *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*, 2nd ed.; Lawrence Erlbaum Associates Publishers, 2007.
(50) Rocabado, G. A.; Kilpatrick, N. A.; Mooring, S. R.; Lewis, J. E. Can We Compare Attitude Scores among Diverse Populations? *J. Chem. Educ.* **2019**, *96*, 2371.
(51) van Sonderen, E.; Sanderman, R.; Coyne, J. C. Ineffectiveness of Reverse Wording of Questionnaire Items: Let's Learn from Cows in the Rain. *PLoS One* **2013**, *8* (7), e68967.
(52) Simpson, R. D.; Rentz, R. R.; Shrum, J. W. Influence of Instrument Characteristics on Student Responses in Attitude Assessment. *J. Res. Sci. Teach.* **1976**, *13* (3), 275−281.