Multilevel Rasch modeling of two-tier multiple choice test: A case study using Lawson's classroom test of scientific reasoning

Yang Xiao,^{1,2} Jing Han,² Kathleen Koenig,³ Jianwen Xiong,^{1,†} and Lei Bao^{2,*}

¹South China Normal University, Guangzhou, Guangdong 510006, China

²The Ohio State University, Columbus, Ohio 43210, USA

³University of Cincinnati, Cincinnati, Ohio 45220, USA

(Received 18 October 2017; published 19 July 2018)

Assessment instruments composed of two-tier multiple choice (TTMC) items are widely used in science education as an effective method to evaluate students' sophisticated understanding. In practice, however, there are often concerns regarding the common scoring methods of TTMC items, which include pair scoring and individual scoring schemes. The pair-scoring method is effective in suppressing "false positives" at the cost of missing possible middle states of progression of student understanding. On the other hand, the individual scoring method captures an undistinguished middle level but is prone to rewarding guessing, which leads to "false positives". In addition, this middle level does not discriminate the progression between knowing the result and explaining the reason, which limits the capacity of drawing meaningful implications from the assessment outcomes. To address the concerns with the current scoring methods, it is valuable to explore new scoring method(s) that can fully utilize the information measured with TTMC items. In this study, a number of scoring models are studied using Rasch analysis on data of a popular TTMC test, the Lawson classroom test of scientific reasoning (LCTSR), collected from four considerably different populations. The results show that the model fit quality of the scoring methods varies with student population and item design. In general, there is no one-fits-all solution; however, given the new information obtained in this study, a three-step process is suggested that can guide the development of new mixed scoring models tailored for a particular population and or test. The evaluation results show that the mixed models produce the most reliable model fitting and better than average goodness of fit. Furthermore, the results in this study also confirm previous studies, which suggest that it is harder to come up with a correct explanation than to just know the answer.

DOI: [10.1103/PhysRevPhysEducRes.14.020104](https://doi.org/10.1103/PhysRevPhysEducRes.14.020104)

I. INTRODUCTION

The current science education standards, i.e., the Next Generation Science Standards (NGSS) and the related framework, call for students to develop coherent understanding of complex science [1,2]. Students with complex understanding should know science concepts, e.g., constructing concepts and disciplinary core ideas, along with scientific practice skills [1,3]. Assessment instruments for tracking students' progression towards sophisticated understanding play a critical role in science education [4]. However, many traditional multiple-choice (MC) tests are

designed primarily for content assessment rather than for measuring higher order cognitive abilities in reasoning and explanation [5,6]. Typical MC items are often critiqued for assessing the superficial memorization of science facts or for the application of simple process skills due to the lack of opportunity for students to explain or justify their answers [7–10], although some research-based instruments, such as the Force Concept Inventory [11], had partially addressed these disadvantages. To fully capture more complex science understanding and reasoning, constructed response items are often used, but the implementation requires a large amount of resources including time and scoring cost and has the weakness of low reliabilities due to human raters [12,13].

To improve the assessment features of MC items, researchers have explored different designs including two-tier multiple choice (TTMC) [14], ordered multiple choice (OMC) [15], multiple-tier multiple choice (3 or more tiers) [16–20], and test combination analysis [21]. This study will focus on the TTMC design, which was first proposed by Treagust [14] and has gained popularity in recent years [12,22–24]. TTMC items are constructed to

*Corresponding author.
bao.15@osu.edu

†Corresponding author.
jwxiong@scnu.edu.cn

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/). Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

jointly measure students' ability to correctly answer tier 1 questions and their explanatory reasoning with tier 2 questions [14,25]. Probing reasoning and explanation in tier 2 provides a way to assess more complex understanding and can also serve to diagnose misconceptions [14]. In recent studies, researchers have made substantial progress in developing TTMC items as effective diagnostic tests [12,22,24,26–31].

In current research, the scoring of a TTMC item includes two popular methods. One is the pair-scoring method, which treats the two questions in a pair as a single item and assigns credit when both are correct. The other is the individual-scoring method, which treats the two questions as two individual items and awards credit for each item independently. Both methods have benefits and weaknesses. For example, the pair-scoring schema suppresses false positives due to guessing as it is less likely for students to guess correctly on two questions. However, this method ignores the possible middle stages of student learning from simple facts to deeper scientific understanding and reasoning. On the other hand, using the individual scoring method awards students for their partially developed correct thinking or memorization during their transition from simple to complex understanding, but this method does not have a mechanism to suppress guessing.

It is therefore beneficial to study how the intermediate learning stages can be more accurately assessed with the TTMC designs while still keeping the effect of guessing suppressed. Towards this assessment goal, there is limited research in the current literature. This warrants new research on scoring methods that can fully utilize the information measured with TTMC items. In this study, a number of scoring methods on TTMC items have been examined using a Rasch model to explore possible scoring designs that may improve assessment using TTMC instruments. The data used in this research were collected using the Lawson classroom test of scientific reasoning (LCTSR), which is a popular TTMC instrument designed for assessing student scientific reasoning skills [32,33].

II. REVIEW OF TTMC ASSESSMENTS

TTMC designs have been widely used in science, technology, engineering and mathematics (STEM) assessment. Examples include misconception identification in science questionnaire [34], the test of image formation by optical reflection [35], student understanding of light and its properties [36], light propagation diagnostic instrument (LPDI) [27], thermodynamics diagnostic test [37], secondary students' scientific reasoning in genetics [24], and the LCTSR [32].

Typically, a TTMC pair is designed to measure a content response in the first tier and the related reason or explanation in the second tier [14]. Therefore, two latent traits are often assumed [14,38]: one for knowing the result and one for explaining the reason. Recently, an empirical study

conducted by Fulmer *et al.* [38] examined this assumption using two TTMC instruments, one on science concepts (light propagation diagnostic Instrument) and the other on scientific reasoning (LCTSR) [27,32]. In their study, items were classified by different tiers representing different latent traits or based on different content subtopics within the instrument, or according to both tier and content. Rasch model analysis for both TTMC instruments showed that the students and items could be distinguished by both content dimensions and the tier of the item, which verified the assumption.

A. Knowing and explaining

The two latent traits embedded in TTMC items can be generally mapped to student abilities regarding knowing the result and explaining the reason, which will be referred to as “knowing and explaining” hereafter. Researchers have always been interested in the progression of ability associated with the two questions in a TTMC pair. Typical research questions include whether students progress from knowing to explaining or from explaining to knowing, and how such progression may be affected by the items and the population. Since the outcome of any assessment is the result of interactions between students and the test items, it can be assumed that the possible direction of progression may have significant dependence on item designs and the population and may vary with these factors. Therefore, the actual patterns of progression for specific test items and population must be determined through empirical studies.

A number of studies using classical test theory (CTT) have provided evidence showing that explaining is more difficult than knowing, which means in certain circumstances students may know the correct answer before having the correct explanation [11,17,18,39–41]. For example, in a study which assessed high school students' understanding of chemistry, significantly more students answered correctly on the first-tier questions (50.1%) than on the second-tier questions (30.0%) [42]. Similar results were found in studies using item response theory (IRT) models [12,38]. For instance, Liu *et al.* found that the second tier explanation items are significantly more difficult than the first tier items on knowing the results [12].

On the other hand, Fulmer *et al.* [38] showed inconsistencies of progression patterns for different types of TTMC tests. For the Light Propagation Diagnostic Instrument (LPDI) [27], tier-2 explanation questions were found to be more difficult than tier-1 content questions. For the LCTSR, the results did not show a consistent pattern by tier or by content sub-topic. Hence, Fulmer *et al.* [38] suggested that two-tier items may not follow any strict pattern in terms of difficulty. Although it has not been extensively discussed in the literature, the origins of inconsistency can come from several areas including the design of the items and the features of the population. In particular, the designs of specific TTMC instruments can

significantly impact the assessment features on knowing and explaining. It is possible that, in some cases, the second-tier items can be used by students as additional information to figure out or confirm their answers to the first question rather than a direct explanation. The inconsistency of the two-tier difficulty discovered in LCTSR by Fulmer *et al.* [38] may also be the result of inconsistencies in item designs. Therefore, this study will examine how student responses to TTMC items vary with the population or the instrument, as well as with different items within the same instrument. Understanding these interactions provides the needed cognitive basis for developing proper scoring models of TTMC items as well as the validity of the two-tier designs regarding knowing and explaining.

B. Scoring methods of TTMC items

There are a number of scoring and analysis methods of TTMC items commonly used in the literature, which include cross tables and numerical scores. The cross tables report the percentages of options being chosen by students across the two tiers of an individual item pair [14,27,40]. Such descriptive analysis is useful in identifying response patterns of possible misconceptions that students may have [14,39,43]; however, the results of the analysis are presented in large cross tables, which are difficult to compare between populations to draw clear-cut conclusions [44]. To address these issues, researchers often assign numerical scores for the two-tiered items, which are then aggregated to produce population measures for advanced psychometric analysis.

As introduced earlier, the assignment of numerical scores to TTMC items includes two popular methods: pair scoring and individual scoring. Pair scoring treats an item pair as a combined item in a dichotomous mode that awards credit for answering both items correctly and zero points for all other responses [26,39,45,46]. For example, the chemistry concept inventory, a widely used assessment instrument, contains six pairs of two-tiered items (Item pair 7-8, 10-11, 12-13, 16-17, 18-19, and 20-21). In a study on psychometric analysis of the instrument based on the Rasch model, Barbera [47] used the pair-scoring schema to score the six item pairs. The results showed that one-third of the two-tier items had outfit values outside the recommended range, which indicates that the scoring schema did not appropriately capture the students' performance and their implied abilities.

The other popular approach uses individual scoring that treats questions in an item pair as individual items and assigns points for each tier independently [12,38,44, 48–52]. As discussed previously, there are both benefits and weaknesses for the two scoring methods. The pair scoring reduces the impact of guessing but also misses the information of possible intermediate stage development, while the individual-scoring awards intermediate levels but is indistinguishable from guessing. Since the two tiers of an item pair are both related as well as distinguished, it would

be useful to fully explore the possible scoring strategies in an attempt to provide better assessment for both the accuracy and the richness of student thinking.

C. Modeling of TTMC responses

To obtain more accurate assessment of students' intermediate level ability measured by TTMC items, more sophisticated models have been explored using item response theory (IRT). One approach involves individual-scoring. For example, in Liu *et al.*'s [12] study the two tiers were graded individually to examine the relationship between the first- and second-tier responses. Similar work was conducted by Fulmer *et al.* [38] as a means of comparing the difficulty levels of the two tiers. However, these studies ignored the strict local independence assumptions of standard IRT models. In fact, applying individual scoring to two-tier items will always violate the local independence assumption, which raises concerns on the validities of the related studies.

To address the local independence issue of the individual-scoring models, a pair of two-tier items can be treated as a single super item and analyzed with polytomous scoring models including the partial credit model, graded response model, and generalized partial credit model [53–55]. In these pairwise polytomous scoring models, multilevel partial credits are assigned to different students' response patterns of a TTMC pair, which include four types of patterns: (i) pattern "00" for incorrect on both result and explanation, (ii) pattern "01" for incorrect result with correct explanation, (iii) pattern "10" for correct result with incorrect explanation, and (iv) pattern "11" for correct on both result and explanation.

The actual assignments of partial credit to the different response patterns would vary with researchers' assumptions on students' learning and ability development. Typically, patterns 00 and 11 would represent the lowest and highest boundaries of the range of partial credit. The relative levels of the two partially correct patterns 10 and 01 are often subject to open research and exploration. It can be argued that the cognitive interpretations of the two partially correct patterns depend on both the item designs and the relation between knowing and explaining [10,56,57]. For example, if student ability is assumed to progress from knowing to explaining, students who answer with incorrect result but correct explanation (pattern 01) should not exist on the progression path and this pattern can only be produced by guessing. On the other hand, students answering with correct result and incorrect explanation (pattern 10) should be considered at a higher level of ability than those who are guessing. As a result, under this assumption, response patterns 00 and 01 can be treated as the lowest level of development while pattern 10 can be considered one level higher than both 00 and 01.

To compare the performance of the different TTMC scoring models, Tam *et al.* [44] conducted a study to

evaluate the model fit of four IRT models [35] including (i) a dichotomous model using individual scoring, (ii) a dichotomous model using pair scoring, (iii) a partial credit model, which assigns 2 points for the pattern 11, 1 point for pattern 10, and 0 points for patterns 00 and 01, and (iv) a partial credit model, which assigns 2 points for the pattern 11, 1 point for pattern 01, and 0 points for patterns 00 and 10. Results of this study suggested that the second dichotomous model outperformed the other three models in model fit. However, this study also had limitations and should not be over generalized: only three two-tier item pairs were used in the measurement, which are too few to obtain adequate item variations. In addition, almost 95% of the students answered with consistent response patterns 00 or 11 such that the data has very limited coverage on the intermediate level patterns 10 and 01. Because of the limitation of the data used in this study, the capacity to evaluate scoring methods that target the inconsistent response patterns is also limited.

D. Research questions

The current literature on assessments that use TTMC covers a broad area of topics from the cognitive strand of knowing and explaining to the various scoring methods and analytic models. In general, assessment variables such as estimated student ability, item difficulty, and other item parameters are interactively related. Students' response patterns on specific item pairs would generally vary with both the population and the item designs. It can be argued that the performance of scoring method should be dependent on test item and population, which to some extent has been implied in previous studies [40,44].

It is therefore important to determine the extent to which features of item and population may influence the performance of assessment models of TTMC items. In this study, Rasch-based TTMC scoring models were examined to explore their item and population dependency. Six Rasch models were compared using data sets of LCTSR taken from four different populations including U.S. high school students, U.S. college students, Chinese middle school students, and Chinese high school students. To comply with the local independence assumption, pairwise partial credit methods were used in the calculation. This study aims to answer three research questions:

1. Are the performances of Rasch-based TTMC scoring models dependent on population and or item design?
2. Can mixed scoring models that apply different TTMC scoring methods for different items improve model performance?
3. Based on the analysis of the different scoring models, is there evidence to support whether explaining reasoning is more advanced than knowing the result or vice versa for the LCTSR and is the result dependent on population and item design?

III. METHODOLOGY

A. Instrument and participants

Data for this study were collected in 2008 and used in a previous study that compared Chinese and U.S. students' scientific reasoning skills [48] using the LCTSR (2000 version) [32,33]. The data set includes four population samples. S1 includes 1953 Chinese middle school students and S2 contains 3409 Chinese high school students. S3 and S4 are comprised of 782 U.S. high school students and 1717 U.S. college students, respectively. These four samples span a broad range of grades and backgrounds, which provide the needed diverse data set to study the possible influence of student population on the scoring models.

The LCTSR consists of 12 two-tier pairs (24 items) [33]. The first 10 pairs (items 1–20) are all typical two-tier design, in which the first tier asks for an outcome and the second tier asks for the reasoning. Item pairs 21–22 and 23–24 were not used in this study as these items were designed to assess hypothetical-deductive reasoning and did not follow the result-explanation structure of the first 10 pairs.

B. Response pattern analysis

To examine the dependency of response patterns on item and population, descriptive cross table analysis was used. The results of response patterns were listed in a cross table spanned with item pairs and populations to examine the possible relations. Cross table and Cochran-Mantel-Haensze chi-square (CMH χ^2) tests [58] were used to determine statistical significance among item pairs, student populations and response patterns. Item pairs and response patterns were treated as two nominal variables in the two-way table. The four populations were treated as four repeats. Hence, the null hypothesis was that the relative proportions of response patterns are independent of the other variable, i.e., item pairs, within the repeats; in other words, there is no consistent difference in proportions in the two-way tables. The alternative hypothesis was that the proportions of response patterns are different for item pairs. For the purpose of verifying the hypothesis in this study, the CMH χ^2 was performed in R following the process suggested by Mangiafico [59].

C. Scoring methods

Using the four response patterns of two-tier item pairs, a range of scoring methods can be developed. In this study, it is assumed that pattern 00 represents the lowest level and pattern 11 represents the highest. This generates a total of six scoring methods (see Table I), which are analyzed with the Rasch model. Existing research has shown that even with an individual-scoring method, which violates the assumption of local independence required by Rasch, the LCTSR agrees well with the unidimensionality assumption and can be analyzed with a Rasch model [49,60]. The

TABLE I. Summary of six scoring methods of two-tier item pairs.

TTMC scoring models	Response pattern and score assignment			
	“00”	“01”	“10”	“11”
M1	0	0	0	1
M2	0	1	1	2
M3	0	1	2	3
M4	0	2	1	3
M5	0	0	1	2
M6	0	1	0	2

approach used in this study will comply more strictly with the assumption of local independence by treating each two-tier item pair as a single super item, which is scored with one of the six scoring methods summarized in Table I.

M1 is the pair-scoring method. M2 will produce the same score as the independent-scoring method, but it is a pairwise method that complies with the requirement of local independence.

M3 and M4 treat the four response patterns as four different levels. M3 assumes that knowing is harder than explaining and there is no guessing of correct answers. On the other hand, M4 follows the assumption that explaining is harder than knowing without considering guessing of the correct explanation.

M5 and M6 both assign three levels for the response patterns. The underlying assumptions are more complicated. M5 treats the pattern 01 as guessing and considers pattern 10 as a meaningful higher ability. The assumption for M5 is that explaining is harder than knowing and providing correct explanation with an incorrect answer implies guessing. Meanwhile, M6 is the opposite of M5 and assumes that knowing is harder than explaining and that providing a correct answer without correct explanation implies guessing.

One of the research questions of this paper is to investigate the relationship between knowing and explaining. This question could be answered by evaluating and comparing among the four models from M3 to M6. If any of the models was found to fit better across population and items, the assumption underlying the corresponding model could then be considered as the supported relation.

D. Rasch model analysis

1. General model

The Rasch model [61] was used in this study to evaluate and compare the performances of the six scoring methods. In Rasch measurement, the ability of the person and item difficulty are put on a common logit scale expressed in Rasch model equations. A dichotomous equation is given as follows:

$$P(X_{ni} = 1) = \frac{e^{B_n - D_i}}{1 + e^{B_n - D_i}}. \quad (1)$$

Here, D_i is the difficulty of item i , B_n is the ability of person n , and X_{ni} is the score of person n on item i . Then the probability for person n to correctly answer item i , $P(X_{ni} = 1)$, is a function of the difference between person ability and item difficulty. If the difference is zero, the probability is 50%. As the difference increases in the positive direction, the person has a higher probability of answering the item correctly.

For polytomous items, the partial credit model (PCM) [53] was used. The PCM model shown in Eq. (2) gives the probability for person n with ability B_n to score $X_{ni} = k$ on item i . Here, D_{ik} is an item difficulty parameter governing the probability of scoring $X_{ni} = k$ rather than $X_{ni} = k - 1$, and m_i is the highest score of item i :

$$P(X_{ni} = k) = \frac{\exp \sum_{v=0}^v (B_n - D_{ik})}{\sum_{v=0}^{m_i} \exp \sum_{k=0}^v (B_n - D_{ik})}, \quad k=0,1,2,\dots,m_i. \quad (2)$$

Among the six scoring methods, M1 is a dichotomous model, while M2–M6 are polytomous models. Measures of person ability and item difficulty from Rasch modeling were used to compare the performances of the six scoring methods over different test items and populations. Based on the Rasch modeling results on person ability and item difficulty, one could determine whether different response patterns belonged to similar or different performance levels. For example, if analysis indicated significant differences between students' average abilities of two response patterns, they would be mapped to two performance levels. Otherwise, the response patterns would be grouped into one performance level.

In addition, the models may behave differently over changing populations and items. To evaluate and compare the changing performances of the different models, the six Rasch models were applied on data sets from all four populations using the TAM R-package [62]. In the analysis, the reliability coefficient and item-level fitting indices were used as evidence to compare the models' goodness of fit across different populations and items. An additional commercial software, Winsteps [63], was used to calculate item reliability, since the TAM package could only produce person reliability.

2. Reliability

In Rasch modeling, the reliability of discriminating students based on their abilities is evaluated with the “separation reliability” coefficient, which shows how consistently the estimated students' abilities match the observed data. This is referred to as person reliability. The coefficient can then be interpreted similarly to Cronbach's alpha in

classical statistics. The separation reliability can also be calculated with item difficulty to evaluate how consistently the model can differentiate the items based on their difficulty, which is referred to as item reliability.

Coefficients of person reliability were used in this study as the evidence for determining the most reliable model(s) under different conditions. The range of a reliability coefficient is from 0 to 1. It is common practice to accept a reliability coefficient of 0.65 or greater as the criteria for reliably differentiating person ability [64]. Accordingly, 0.65 was set as the acceptable criteria and a higher reliability coefficient was considered better. In this study, coefficients of item reliability of the different models were all very close to 1 due to the large sample size, and therefore were not used in the analysis.

3. Goodness of model fit

Mean-square (MNSQ) infit and outfit indices were used to evaluate the goodness of model fit for different items, while the standardized fit statistics (ZSTD) indices were not adopted in this study because of the large sample size [65]. The MNSQ fit indices were calculated based on the differences between observed and expected responses. Infit MNSQ refers to inlier-sensitive or information-weighted fit, which is more sensitive to responses to items with difficulty targeted on the person, and vice versa. Meanwhile, outfit MNSQ refers to outlier-sensitive fit. This is more sensitive to responses to items with difficulty far from a person, and vice versa.

A fundamental assumption of using the Rasch model is that if an item is an effective measure of students' abilities, then the probability of a student answering the item correctly should increase monotonically with the student's ability. MNSQ fit statistics allow the evaluation of the conformity of data to this assumption. For a well-fitted model, MNSQ fit indices have an expected value of 1.0, and values between 0.7 and 1.3 suggest a good fit between model and items [61]. Both infit and outfit indices were used to compare the models' fitting performances at both the instrument and item level. The closer the mean-square fit indices approach the expected value of 1.0, the better the model fits.

Furthermore, the average goodness of model fits of multiple items was evaluated based on the root mean square error (RMSE) of infit and outfit measures of all items in an instrument. RMSE is commonly employed in model evaluation studies [66]. RMSE gives the root-mean-square value of the differences between an observed fit index and the expected best fit value of 1.0. The actual calculation is shown as follows:

$$\text{RMSE} = \sqrt{\frac{1}{m} \sum_{j=1}^m (y_j - 1.0)^2}. \quad (3)$$

Here, y_j is the observed fit index of the j th item pair and m is the total number of item pairs. A lower value of RMSE indicates an average better fit of the model. Note that the RMSE can be interpreted as the standard deviation of the fit indices.

4. Developing mixed scoring models

For the four response patterns, the average student abilities were calculated and compared. When the average abilities of different response patterns were significantly different from one another, the patterns were distinguished as representing different ability levels. Otherwise, patterns with similar abilities were grouped into a single ability level. The statistical significance was evaluated with a two-way ANOVA to examine how student ability varied across patterns and item pairs.

To further examine whether or not it is valid to assign different performance levels to students' response patterns, item response curves (IRCs) were also used to evaluate if the distributions of such curves warrant different performance characteristics. IRCs are similar to the category probability curves (CPCs) from the IRT framework [67,68]. CPCs are visual representations of the probabilistic relationship between category response and student ability on the logit scale based on the item or category response function. Similarly, IRCs also describe the probability distribution of students' answering with a particular response pattern as a function of students' ability. The difference is that the probabilities of IRCs are directly calculated from students' response data rather than predicted probabilities from model fitting.

The main reason for using IRCs instead of CPCs is that CPCs of specific response patterns have to be obtained by using the patterns in the scoring models, which are different from one another and may not include all of the patterns. If a pattern is not used as a unique level for scoring (i.e., it is grouped with other patterns), the CPC of this pattern cannot be directly obtained through the modeling output. Using IRCs avoids this issue and IRCs of all patterns can be obtained using students' abilities estimated with any model.

In addition, since different models use different scoring methods, their estimated students' abilities do not have a common scoring base, and therefore cannot be directly compared. To address this issue, all students' abilities were estimated using M1 so that these could be directly compared and IRCs were used to show the probability distributions of the four response patterns for all item pairs and populations.

Distributions of IRCs can provide detailed information on how the corresponding response patterns discriminate students of different abilities. Ideally, students of lower ability are more likely to answer with the lower-level patterns and students of higher ability are more likely to answer with the higher-level patterns. When the probability

TABLE II. Statistics of students' response patterns to items 1–20 of the LCTSR.

Samples	Patterns	Mean	1-2	3-4	Percentages of response patterns (%)							
					5-6	7-8	9-10	11-12	13-14	15-16	17-18	19-20
S1	00	42.30	14.34	32.26	35.33	37.12	37.99	69.64	40.19	48.44	58.47	49.26
	01	8.51	1.59	1.18	10.70	12.75	6.09	9.11	15.21	14.34	7.32	6.81
	10	13.15	4.66	5.02	8.09	20.74	21.15	16.23	26.52	7.42	7.68	14.03
	11	36.03	79.42	61.55	45.88	29.39	34.77	5.02	18.07	29.80	26.52	29.90
	Sum of 01 and 10	21.67	6.25	6.20	18.79	33.49	27.24	25.35	41.73	21.76	15.00	20.84
	Difficulty	0.87	−1.70	−0.60	0.21	1.11	0.80	3.54	1.90	1.09	1.29	1.08
	SE		0.06	0.05	0.05	0.06	0.05	0.11	0.06	0.06	0.06	0.06
	Pt-bis		0.46	0.51	0.65	0.56	0.58	0.23	0.37	0.55	0.56	0.50
S2	00	27.96	6.42	12.23	17.72	20.21	15.84	53.89	29.30	36.61	47.46	39.95
	01	6.51	1.17	1.91	7.30	9.62	6.60	6.69	12.67	9.59	4.40	5.13
	10	12.73	3.55	3.23	6.51	24.70	16.10	23.82	23.53	6.72	7.45	11.67
	11	52.80	88.85	82.63	68.47	45.47	61.45	15.61	34.50	47.08	40.69	43.24
	Sum of 01 and 10	19.24	4.72	5.13	13.82	34.32	22.70	30.51	36.20	16.31	11.85	16.81
	Difficulty	−0.188	−2.49	−1.89	−0.95	0.23	−0.57	2.03	0.80	0.15	0.47	0.34
	SE		0.06	0.05	0.04	0.04	0.04	0.05	0.04	0.04	0.04	0.04
	Pt-bis		0.44	0.52	0.59	0.48	0.51	0.35	0.44	0.56	0.55	0.48
S3	00	31.05	6.01	33.25	55.88	41.43	28.52	53.84	34.78	9.34	18.93	28.52
	01	8.57	1.28	0.64	5.75	31.71	3.20	8.06	14.07	14.58	3.32	3.07
	10	8.84	2.69	1.02	5.12	2.81	6.01	23.40	24.55	2.30	5.63	14.83
	11	51.55	90.03	65.09	33.25	24.04	62.28	14.71	26.60	73.79	72.12	53.58
	Sum of 01 and 10	17.40	3.96	1.66	10.87	34.53	9.21	31.46	38.62	16.88	8.95	17.90
	Difficulty	−0.166	−3.06	−0.91	1.02	1.66	−0.74	2.49	1.47	−1.50	−1.38	−0.21
	SE		0.13	0.09	0.09	0.10	0.09	0.12	0.10	0.10	0.10	0.09
	Pt-bis		0.34	0.52	0.65	0.59	0.65	0.41	0.54	0.65	0.65	0.64
S4	00	20.51	2.80	20.85	31.04	23.18	18.87	41.47	24.81	7.11	12.87	22.07
	01	7.56	1.28	0.41	4.48	27.26	1.98	7.11	16.07	10.60	2.97	3.44
	10	8.25	1.57	1.05	4.54	6.58	2.80	27.49	23.30	1.22	3.79	10.19
	11	63.68	94.35	77.69	59.93	42.98	76.35	23.94	35.82	81.07	80.37	64.30
	Sum of 01 and 10	17.40	3.96	1.46	9.03	33.84	4.78	34.60	39.37	11.82	6.76	13.63
	Difficulty	−0.93	−3.61	−1.68	−0.54	0.40	−1.58	1.57	0.81	−1.95	−1.89	−0.79
	SE		0.11	0.07	0.06	0.06	0.07	0.06	0.06	0.07	0.07	0.06
	Pt-bis		0.33	0.53	0.67	0.62	0.55	0.45	0.54	0.59	0.59	0.52

distribution curves show different peaks and distribution shapes, it is appropriate to classify the corresponding response patterns into multiple levels. On the other hand, when the probability curves overlap each other, the response patterns represented by these curves should be grouped into one ability level.

IV. RESULTS

A. Dependency of response patterns on item design and population

Descriptive statistics of students' response patterns are given in Table II, which summarizes the mean percentages of the four response patterns for different items and with different populations. The consistent patterns 00 and 11 typically have straightforward relations with the item difficulties and student ability. The mixed patterns 10

and 01, however, may have varied levels of implied abilities that also vary with item and population.

The results in Table II suggest that both the consistent and mixed patterns vary substantially over population and item. For example, comparing item pairs 7-8 and 15-16 for the S2 population shows that they both have similar 11 patterns (45.47% vs 47.08%) which indicates a similar difficulty level for this population. However, their mixed patterns are significantly different, where pair 7-8 has a distinctively larger 10 pattern (24.70%) than that of the pair 15-16 (6.72%). When comparing with S4, the consistent and mixed patterns change substantially from that of S2. For S4, pair 15-16 has a much larger 11 pattern (81.07%) than pair 7-8 does (42.98%) suggesting that pair 15-16 is much easier for this population. Meanwhile, for the mixed patterns, pair 7-8 has a larger 01 pattern (27.26%) than the 10 pattern for S4, showing that more students perform

correctly on explaining than on knowing. This is opposite to the common assumption on the progression from knowing to explaining and may indicate possible issues in item design.

To evaluate the statistical significance of the variations of response patterns with respect to changes in item and population, the Cochran-Mantel-Haenszel chi-square (CMH χ^2) test was applied on the data in Table II. The analysis shows that the variations are statistically significant ($\chi^2_{MH} = 557.57$, $df = 27$, $p < 0.001$), indicating that students' response patterns are dependent on both item and population.

In addition to descriptive statistics, Rasch analysis parameters using M1 were also given in Table II, which included Rasch difficulty in logit, Rasch difficulty standard error, and point biserial correlation. Infit and outfit MNSQ statistics were not displayed here as these will be used to evaluate the goodness of model fit in later sections. As shown in Table II, the item difficulties are well distributed for all populations. The Rasch difficulty standard error, which gives the uncertainty of an estimated difficulty, is in the range of 0.06–0.13 showing good reliability. The point biserial correlation is a measure of consistency between a single item and the whole test. The results show that all items in the LCTSR satisfy the criterion of the point biserial coefficient (Pt-bis ≥ 0.2) as suggested by Kline [69].

In Rasch modeling, the evaluation of instrument quality is often analyzed with Wright maps, which compare the distributions of item difficulty and person ability. A Wright map provides information on how well persons and items are distributed along the ability-difficulty logit scale. Wright maps of LCTSR for all four populations modeled with M1 are given in the Supplemental Material [70]. The results show that the LCTSR items distribute well along a good range of person ability scale.

B. Model evaluation and optimization

1. Reliability

The reliability of the different model fits was evaluated using person reliability coefficients, which are summarized in Table III for all models and populations. For the discussion in this section, results of M1 to M6 are used.

Also shown in the table is the MSM model (mixed scoring model), which is a new model that uses mixed scoring methods. This model will be introduced later and its results should be ignored until then.

According to DeVellis [64], a reliability of 0.65–0.70 is “minimally acceptable” and a reliability between 0.70 and 0.85 is “respectable” for instruments to be used for research purposes. The results in Table III suggest that none of the six models are overwhelmingly superior to other models and their fitting reliabilities vary substantially with the population. That is, there is no one model that has the most reliable fit for all populations. Meanwhile, certain populations may fit several models more reliably than other populations.

Overall, S3 seems to give the most reliable fit for all models while S2 shows the lowest reliability. Each population also has its own best-fitting models regarding item and personal reliability, which vary from population to population. For example, M5 is the best fit for S2. This suggests that it would be more reliable to score this population by classifying the four response patterns into three categories as defined in Table I for M5 (00-01, 10, 11).

Apparently, the reliability of model fit is dependent on population. In addition, as indicated in Table III, the different scoring models perform differently within a single population. This implies that model fitting performance also depends on the items of an instrument. It is therefore important to fully examine how such dependence manifests through varying populations and item designs. Since S2 shows the weakest fit according to item and person reliabilities in Table III, in the analysis that follows, S2 will be used to investigate item-level model fitting performances.

2. Goodness of model fit

Using data from S2, infit and outfit MNSQ indices of the six models were calculated for each item pair and listed in Table IV. The majority of items on the LCTSR fit well with the Rasch model, which have infit and outfit MNSQ indices close to 1.0 and fall within the “good fit” criteria range of 0.7–1.3. Among all of the items, item pairs 1-2 and 3-4

TABLE III. Model fitting person reliability coefficients. For M1–M6, the highest reliability coefficients of each population are underlined. The reliability coefficients above the “minimally acceptable” level (>0.65) are marked with an asterisk (*) and the reliability coefficients above the “respectable” level (>0.70) are marked with two asterisks (**). MSM, which will be introduced later, is a new Rasch model based on a mixed scoring method. Statistical analysis of the reliability measures is discussed in the Supplemental Material [70].

		M1	M2	M3	M4	M5	M6	MSM
S1	Person reliability	<u>0.674*</u>	0.653*	0.654*	0.643	0.673*	0.641	0.687
S2	Person reliability	<u>0.657*</u>	0.658*	0.642	0.642	<u>0.673*</u>	0.628	0.680
S3	Person reliability	<u>0.774**</u>	0.763**	0.759**	0.751**	<u>0.769**</u>	0.760**	0.777
S4	Person reliability	<u>0.722**</u>	0.720**	0.720**	0.709**	0.724**	<u>0.727**</u>	0.740

TABLE IV. Item fit indices of S2. For M1–M6, the indices outside the range of 0.7–1.3 are marked with an asterisk (*) and the indices closest to 1.0 are underlined. Statistical analysis of the infit and outfit measures is discussed in the Supplemental Material [70].

Item pair	M1	M2	M3	M4	M5	M6	MSM
Infit MNSQ measures							
1–2	0.793	0.781	<u>0.796</u>	0.781	<u>0.796</u>	0.771	0.830
3–4	0.802	0.794	<u>0.805</u>	0.793	<u>0.807</u>	0.787	0.823
5–6	<u>0.832</u>	0.796	0.802	0.804	<u>0.819</u>	0.819	0.785
7–8	<u>0.996</u>	0.918	0.884	0.964	0.917	1.014	0.912
9–10	<u>0.958</u>	0.892	0.880	0.913	0.911	0.945	0.944
11–12	0.942	0.954	0.949	0.970	0.920	<u>0.978</u>	0.939
13–14	0.993	<u>0.994</u>	0.993	1.007	0.993	<u>1.013</u>	1.010
15–16	0.891	<u>0.893</u>	0.891	<u>0.898</u>	0.897	0.895	0.914
17–18	0.888	0.898	<u>0.912</u>	<u>0.885</u>	0.909	0.883	0.917
19–20	0.986	1.067	1.091	1.034	1.066	<u>0.987</u>	0.986
RMSE	0.118	0.134	0.133	0.128	0.125	0.126	0.116
Outfit MNSQ measures							
1–2	0.671*	0.664*	0.691*	0.645*	<u>0.700</u>	0.639*	0.699
3–4	0.675*	0.675*	0.677*	0.674*	<u>0.687*</u>	0.667*	0.700
5–6	<u>0.737</u>	0.705	0.704	0.712	<u>0.726</u>	0.732	0.708
7–8	<u>0.944</u>	0.901	0.842	0.962	0.885	<u>1.004</u>	0.909
9–10	0.889	0.843	0.820	0.866	0.858	<u>0.898</u>	0.871
11–12	0.961	<u>0.986</u>	0.964	1.035	0.944	<u>1.072</u>	1.015
13–14	1.027	<u>0.976</u>	0.966	<u>0.996</u>	0.982	1.013	1.001
15–16	0.844	0.837	0.831	<u>0.848</u>	0.847	<u>0.858</u>	0.878
17–18	0.850	0.886	0.904	0.864	<u>0.910</u>	0.844	0.885
19–20	0.946	1.059	1.085	<u>1.019</u>	1.057	0.962	0.962
RMSE	0.187	0.196	0.198	0.194	0.183	0.194	0.179

seem to produce the worse fit on infit and outfit MNSQ. As expected, the model fit varies with both items and models.

To evaluate the average goodness of model fit, the root mean square error of infit and outfit measures of all items was calculated for the six models as shown in Table IV. Comparing the RMSE infit and outfit results of the different models, M1 and M5 appear to fit slightly better than the other models. However, results in Tables III and IV clearly show a strong dependence of scoring models on population and item design. In general, one should not expect that one model will fit all populations and items. The results also suggest that when selecting a scoring model, researchers may need to evaluate the targeted population with all alternative models and select the best fitting one.

In addition, even when the average best fitting model can be chosen for a specific population, the model in general will not fit all items equally. To address this issue, a new approach is proposed, which suggests the need to develop a type of mixed model that combines different scoring methods such that each fits a subset of the items best. In the second study presented in this paper, a mixed model will be developed and compared with the six models discussed earlier.

C. Developing and testing mixed scoring models

1. Students' abilities represented by different response patterns

The motivation to study the different response patterns is that such patterns may provide measures of student ability at finer grain sizes. As shown by the results discussed earlier, the different scoring models, which are built on the response patterns, are dependent on both the populations and items. As a result, it is unlikely to have a universally superior model that would fit best for all populations and items. On the other hand, for a specific population it is possible to develop a mixed model, which uses not one but a combination of the different scoring patterns that each fits best for a subset of the items. To design such mixed models, it is important to examine in detail how the different response patterns discriminate students on their ability scales. In the following analysis, a number of methods will be used to inspect the discrimination features of the four response patterns on different items and with different populations.

First, the average students' abilities for each response pattern across different populations was compared as shown in Fig. 1. For this analysis, the student ability was calculated with all four populations merged into a super sample so that the ability values were on the same scale and suitable for direct comparisons. In addition, M1 was used as the model to calculate the ability. The reason to use M1 among others is that M1 represents the simplest two-level model in which response patterns 00, 01, and 10 are all grouped into one category with zero points. This provides a baseline configuration without introducing any artificial assumptions of how the mixed patterns 10 and 01 may have differed from 00 and one another.

The results in Fig. 1 show that the ability measures of the different response patterns vary significantly with population. For S1 and S2, although their abilities are quite different, the response patterns behave similarly. Patterns 00 and 01 are nearly identical on the ability scale ($p = 0.783$ and 0.585 , respectively) and 10 is slightly

FIG. 1. Average students' abilities of the four response patterns. The students' abilities are all estimated using M1 with all four populations combined to allow a consistent comparison.

but statistically higher ($p < 0.001$). Here, the p values are corrected following the Bonferroni correction procedure for multiple comparisons among response patterns [71]. For these two populations, the three-level scoring model M5 (00-01, 10, 11) is most appropriate. For S3, the two mixed patterns (10 and 01) are indistinguishable ($p = 0.162$) but are both statistically different from patterns 00 and 11 ($p < 0.001$). As a result, a different three level model M2 (00, 10-01, 11) is suggested. For S4, which has the highest student ability, all four patterns are statistically different from each other ($p < 0.001$) and therefore the four-level model M3 (00, 01, 10, 11) is most suitable.

In addition, among all four populations, a consistent trend can be observed, which shows that the average student ability of pattern 10 is always higher than that of pattern 01. It suggests that for the LCTSR, students who are able to pick the correct answer without correct explanation have higher abilities than students who fail to pick the correct answer but selected the correct explanation. This is consistent with the assumption on the progression of student ability from knowing to explaining [16,17,38–41]. Meanwhile, for those students who pick the correct explanation without knowing the correct answer, it is likely that these students were using guessing strategies or relying on a primitive sense of related information [50]. As a result, these students would have lower abilities in general.

Second, the average ability measures of the four response patterns over different items were calculated

and plotted in Figs. 2(a)–2(d) for the four populations. Here the student ability was calculated using M1 with each individual population to allow for consistent comparisons across different items. In the diagrams, the error bars were not plotted to allow clearer presentation of the details. The standard error is typically less than 0.05, which makes any differences larger than 0.10 statistically significant at a 5% confidence level.

Results in Fig. 2 clearly show that the student abilities of the four response patterns vary significantly across different items and populations. The two-way ANOVA analysis indicates significant interactions between response patterns and items for all four populations, $F_{s1}(27, 19490) = 10.894$, $p < 0.001$, $\eta_p^2 = 0.015$, $F_{s2}(27, 34050) = 23.495$, $p < 0.001$, $\eta_p^2 = 0.019$, $F_{s3}(27, 7780) = 4.627$, $p < 0.001$, $\eta_p^2 = 0.016$, $F_{s4}(27, 17130) = 10.894$, $p < 0.001$, $\eta_p^2 = 0.019$. In addition, the results suggest significant main effects for both response patterns [$F_{s1}(3, 19490) = 1751.095$, $p < 0.001$, $\eta_p^2 = 0.212$, $F_{s2}(3, 34050) = 3548.108$, $p < 0.001$, $\eta_p^2 = 0.238$, $F_{s3}(3, 7780) = 1049.002$, $p < 0.001$, $\eta_p^2 = 0.288$, $F_{s4}(3, 17130) = 1852.823$, $p < 0.001$, $\eta_p^2 = 0.245$] and items [$F_{s1}(9, 19490) = 44.498$, $p < 0.001$, $\eta_p^2 = 0.020$, $F_{s2}(9, 34050) = 119.387$, $p < 0.001$, $\eta_p^2 = 0.031$, $F_{s3}(9, 7780) = 45.961$, $p < 0.001$, $\eta_p^2 = 0.050$, $F_{s4}(9, 17130) = 68.1$, $p < 0.001$, $\eta_p^2 = 0.035$].

FIG. 2. Item-level average abilities of students answering with different response patterns. The students' abilities are estimated using M1 within each population.

When inspecting the results in Fig. 2 at the item pair level interesting details can be revealed regarding how individual item pairs may be scored with different models and how the performances vary with items and populations. For example, for item pair 1-2, the results suggest using M1 for S1 and M6 for S2, S3, and S4. On the other hand, for item pair 9-10, the results suggest M5 for S1 and S2, and M1 for S3 and S4. To accurately determine the statistical significance of the differences between the different response patterns at the item pair level, the mean abilities and their standard errors of all item pairs are listed in Table V for population S2 along with the suggested models and the maximum p values between distinctive response patterns within each model. The average abilities of the four response patterns over all item pairs are also given, which suggest M5 as the best model (see Table V and Fig. 1). M5 is a three-level model treating pattern 01 as guessing and pattern 10 as a meaningful level with higher ability than patterns 00 and 01. This result further indicates that on average explaining is harder than knowing for the LCTSR with S2.

The results and analysis show an operational approach in designing a mixed scoring model that may perform better than any of the single-method models. The item-pair level statistics provide the basis for the initial construction of such a mixed model, which can be further validated and refined through additional analysis of item response characters and model fitting evaluations.

2. Item response curves analysis

To further inspect the validities of the different scoring models suggested in Table V, item response curves (IRCs) are analyzed to examine how students with different abilities may answer using the different response patterns. The shape of the IRC distribution can be used as additional evidence to determine if the choice of a model can help distinguish different types of students. For example, if two response patterns have a similar distribution shape (e.g., one distribution is on top of the other with similar shapes

that only differ in vertical scale), they would represent the same type of person ability characteristics and should be grouped into one scoring level. On the other hand, if two response patterns have different distribution shapes such as with obviously different peak locations or having crossing overs, they would represent different person ability characteristics and should be set as different scoring levels. Figure 3 shows the IRCs of the response patterns for all ten item pairs for S2. Again, the students' abilities were calculated using M1 to allow consistent comparisons over the different item pairs.

From Table V, item pairs 1-2, 3-4, 5-6, 15-16, 17-18, and 19-20 are recommended to be modeled with M1. In Fig. 3, the IRCs of these item pairs show interesting similarities: the IRCs of the mix patterns (01 and 10) have very similar shapes and are also similar to the IRC of pattern 00. Although the magnitude of the IRC of pattern 00 is much larger, the overall distribution shapes are quite similar. This result supports the choice of M1, which groups the three response patterns (00, 01, and 10) into one scoring level of zero points.

For item pairs 7-8, 9-10, 11-12, and 13-14, Table V suggests different multilevel scoring models. Results in Fig. 3 also show supportive evidence for these choices: one or both of the IRCs of the two mixed patterns have very different distribution shapes from the IRCs of patterns 00 and 11. Therefore, these item pairs are justified to use multilevel scoring models. To summarize, the analysis of the IRCs in Fig. 3 provides further evidence that supports the use of the different models recommended in Table V for all item pairs except for item-pair 1-2, which is changed to M1 based on its IRC distributions.

Based on the analysis and discussions, an optimized mixed scoring model (MSM) can be developed through three general steps: (i) the average students' abilities of different response patterns need to be compared, which provide the basic incentive to explore whether or not the different response patterns discriminate students' abilities;

TABLE V. Item-level students' abilities of response patterns of S2.

Item pair	Pattern 00		Pattern 01		Pattern 10		Pattern 11		Model (scoring)	p
	Mean	SE	Mean	SE	Mean	SE	Mean	SE		
1-2	-1.848	0.104	-1.335	0.208	-1.918	0.143	0.199	0.021	M1 (0001)	<0.001
3-4	-1.623	0.069	-1.397	0.153	-1.609	0.134	0.304	0.020	M1 (0001)	<0.001
5-6	-1.277	0.050	-0.965	0.086	-1.093	0.085	0.500	0.021	M1 (0001)	0.002
7-8	-1.114	0.048	-0.859	0.071	-0.082	0.039	0.665	0.026	M3 (0123)	0.003
9-10	-1.198	0.057	-1.059	0.101	-0.505	0.048	0.513	0.023	M5 (0012)	<0.001
11-12	-0.397	0.028	-0.581	0.093	0.248	0.043	1.075	0.053	M5 (0012)	<0.001
13-14	-0.644	0.042	-0.360	0.053	-0.250	0.043	0.776	0.033	M2 (0112)	<0.001
15-16	-0.697	0.033	-0.683	0.064	-0.789	0.091	0.739	0.026	M1 (0001)	<0.001
17-18	-0.591	0.029	-0.627	0.102	-0.711	0.091	0.825	0.028	M1 (0001)	<0.001
19-20	-0.495	0.034	-0.659	0.097	-0.815	0.066	0.696	0.028	M1 (0001)	<0.001
Total mean	-0.749	0.013	-0.732	0.028	-0.388	0.020	0.532	0.008	M5 (0012)	<0.001

FIG. 3. Item response curves (IRCs) of S2. Student ability was estimated based on M1.

FIG. 4. Item response curves (IRCs) of S2 obtained with the new mixed Rasch model.

(ii) the average abilities of all response patterns over different items are then calculated and plotted for each population. This provides detailed information on whether item-level scoring optimization is needed for a specific population. When needed, statistical analysis must be applied to determine the significance of the discrimination on students' abilities by the different response patterns. The results are then used to suggest the most appropriate models for individual item pairs; (iii) IRCs of all item pairs are analyzed to provide further evidence on the validities of selecting the different scoring models suggested in the second step. If needed, adjustment may be made to produce the final structure of the optimal model.

3. Evaluation of the mixed scoring models

Based on the statistical evaluations and IRC analysis, an optimal mixed scoring model (MSM) was obtained, which agrees with Table V. In this section, the MSM will be applied to S2 using the partial credit model [53]. The results will be compared with the other six scoring models to evaluate the fitting performance of MSM.

The person reliability coefficient of the MSM is 0.680, which is the highest among all models for S2 (see Table III). This result indicates that using the new model can improve the fitting reliability. A possible explanation for the increase in reliability is that when each item pair is scored with its optimal model, students' abilities can be estimated more precisely and produce a more accurate overall logit scale for the population. This will increase the agreement in sample-item alignment and thus improve the reliability.

Regarding the item-level MNSQ infit and outfit indices, the new model also has the best overall fit as indicated by the RMSE measures given in Table IV. At the individual item level, the fitting performance varies. This is expected as changing scoring models will completely redefine the scale of the estimated students' abilities, which will impact the fitting of individual item pairs. The effects can be observed through the changes of IRCs of the same item pairs shown in Figs. 3 and 4. The main focus here is the overall fitting performance, on which the new model has consistently demonstrated its improvement in both reliability and average item-level infit and outfit indices.

Comparing the IRCs in Fig. 4 with those in Fig. 3, one can obtain further evidence of improvement by using the new model. In the new model, item pairs 1-2, 3-4, 5-6, 15-16, 17-18, and 19-20 are scored dichotomously (with M1) and item pairs 7-8, 9-10, 11-12, and 13-14 are scored polytomously (with one of M2, M3, and M5). For the dichotomous group of item pairs, IRCs of the low popularity mixed patterns shown in Fig. 3 are now grouped with the 00 pattern. As a result, two clear-cut IRCs are given as shown in Fig. 4. For the polytomous group of item pairs, the mixed patterns also stand out more clearly than the ones in Fig. 3.

Synthesizing the analysis and comparisons, the results of the new model indicate the initial success of developing a mixed scoring model. Using the same process, mixed scoring models (MSMs) for the other three populations were also developed and tested (see Supplemental Material for details). The results show that using the new MSMs generally improve model fitting reliability, which was observed in all three populations (see Table III).

The item-level infit and outfit indices were also calculated (see the Supplemental Material [70]). The average RMSE measures show that the new MSMs provide the best fit on S4 but perform mediocrally on S1 and S3. However, it is encouraging to see that some MNSQ indices that were originally outside of the range of 0.7–1.3 have moved into or closer to the range with the MSMs (e.g., item pairs 15-16 and 17-18 for S3 and item pairs 5-6 and 17-18 for S4).

The overall evaluation of the new MSM approach is encouraging. The MSMs improve reliability for nearly all populations and produce the best fits for some of the populations. The results suggest that even with the new mixed scoring models that fit best for individual item pairs, the overall fitting performance still varies with the population, which further confirms the hypothesis that no single model will fit best for all populations. On the other hand, the new MSMs can be a viable method to search for the “better” model for a range of populations. It is also suggested that for a given population, one may need to go through the evaluation process outlined in this study in order to identify the optimal model. In addition, it is also possible to develop adaptive MSMs that are constructed based on a regression process through model fitting evaluations, which modifies the item-level scoring methods to find the best model. However, the outcome of such a model may not provide a clear assessment meaning on the basis of model construction as it is largely based on mathematical manipulations rather than outcomes from cognitive and empirical studies.

V. SUMMARY AND DISCUSSION

Instruments composed of TTMC items are widely used in science education as an effective method to assess students' sophisticated understanding. In practice, however, researchers have yet to achieve a consensus on TTMC scoring methods. The two traditional scoring methods have both undergone scrutiny regarding their validities and potential weaknesses. The first approach, which gives one point when both tiers are correct, cannot completely capture the intermediate levels of students' progression of understanding. The second approach, which gives one point for each tier independently, ignores the possible relationship between knowing and explaining and has a tendency to produce a “false positive” by giving credit to students who may be guessing. A possible reason for the inconsistency and argument of the validity and performance of the different scoring methods is that these scoring

methods may be population and item dependent. Therefore, if a study only uses a narrowly selected population, its results can be very different from a study of a different population.

Besides these two common scoring methods, a total of six scoring models can be obtained from the four response patterns of a two-tier item pair, assuming that pattern 00 is the lowest level and pattern 11 is the highest. This study evaluated the six models and has demonstrated that the effectiveness of the different scoring models indeed varies with student population and item design. In general, there is no single scoring model that would produce the best fit for all populations and or test items. As a result, researchers need to evaluate all of the related models in order to identify the most appropriate ones for their specific populations. This outcome provides a direct answer to the first research question posed in this paper.

To answer the second research question, this study proposed and tested a method for creating mixed scoring models that use different scoring rules for different TTMC item pairs in a single test. The mixed scoring models were developed through three general steps: (i) the average students' abilities of different response patterns were compared to determine the validity of using response patterns to discriminate students' abilities; (ii) the average abilities of all response patterns over different items were analyzed to suggest the most appropriate models for individual item pairs; and (iii) IRCs of all item pairs were analyzed to further validate the suggested model and to make adjustments, if needed. The evaluation results suggest that the new mixed models may provide better performance than the single-method scoring models for certain populations.

Regarding the third research question, the results in Fig. 1 clearly show that on average for the 10 item pairs of the LCTSR, knowing the correct answer for tier 1 without correct explanation (pattern 10) represents a higher level of student ability than having correct explanation without correct answer (pattern 01). This confirms that the response pattern 01 should in general be considered as the result of guessing. This is consistent with previous studies which suggested that explaining is harder than knowing [11,16,38,40].

At the item pair level, the results vary for different items and are also affected by the population. As shown in Fig. 2, for the majority of the items, pattern 10 has a higher average ability than pattern 01. However, on some of the item pairs (such as item pairs 1-2 and 3-4), the situation is reversed for some populations such as S3 and S4. Again, the results are dependent on both the item design and the population. It is worth clarifying that the results in this study are based on statistics of students' response patterns. Although the response patterns are assumed to represent knowing and explaining, the item pairs may have design issues that might not be able to accurately assess certain populations' actual thinking and reasoning. Nevertheless, based on the results and analysis from this study, the third

research question can be answered in the context of the LCTSR. That is, for this particular test, it appears that explaining is harder than knowing and the actual measurement results are dependent on population and item design, which may further suggest different pathways of development in reasoning ability at different ages and or item design issues that have not been fully validated and determined.

The main goal of this study is to explore the item-population dependency of the different modeling approaches for TTMC instruments and LCTSR is used as an example for this exploration. This establishes the utility of this study for conducting model evaluation in assessment analysis and instrument development. The practical implications of this work include several areas. For the assessment of a single population with an existing instrument, one can evaluate and find the best single-pattern model or create a mixed scoring model to obtain more in-depth and finer grain-sized measures of ability progressions. For comparing multiple populations, one may evaluate all the related models, including mixed scoring models, to find one that has the overall better fit for all populations and use it consistently to generate comparable results.

The actual choice of one model over another depends on both the models' fitting performances and the goals of the research. As discussed in detail in the Supplemental Material [70], with the data of this study, M1 appears to have the best fitting performance. This is intrinsic to the nature of M1 since it involves smallest number of dimensions which will reduce uncertainty and misfit. If one aims to have the most reliable fit with less detail, then M1 would be the best model but it will underestimate the student ability. If one needs to discriminate between "knowing only" and "knowing and explaining," then M5 is the better model at the cost of slightly larger uncertainty as it increased its assessment dimensions. Therefore, with a specific test, a perfect model, which provides the best fitting and discrimination, may not exist and there always needs to be some trade-off in deciding on the model.

In general, if the goal of the research is to obtain the most reliable comparison of students' ability, then the choice of model should be mostly based on model performance. In this study, M1 would be chosen for this purpose. However, if the goal of the research is to distinguish finer levels between knowing and explaining, M5 would be a good candidate. Meanwhile, MSM provides an adaptive approach to balance between reliability and discrimination, and therefore is also a reasonable choice.

For developing and validating an instrument, using the methods developed in this study, researchers can pilot test and evaluate the instrument to explore the population and item dependency of the possible scoring models. If the goal is to use a fixed scoring model for a range of populations, researchers can use the model evaluation method to modify and reevaluate the items until the desired scoring model

demonstrates satisfactory performance for the targeted populations. Based on the insight in the progression from knowing to explaining gained from this study, it can also be suggested that if a single-pattern scoring method is favored, M5, which is consistent with the claim that explaining is harder than knowing, should be considered as a better target model for developing a TTMC instrument. In such a development, researchers should refine the instrument so that M5 can produce the best fit among other scoring models as well as a meaningful discrimination.

ACKNOWLEDGMENTS

The research is supported in part by NSF Grants No. DUE-1044724, No. DUE-1431908, No. DRL-1417983, and No. DUE-1712238, and by the top graduate student scholarship of South China Normal University. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

-
- [1] National Research Council, *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas* (National Academies Press, Washington, DC, 2012).
 - [2] National Research Council, *Next Generation Science Standards: For States, by States* (National Academies Press, Washington, DC, 2013).
 - [3] M. C. Linn, H.-S. Lee, R. Tinker, F. Husic, and J. L. Chiu, Teaching and assessing knowledge integration in science, *Science* **313**, 1049 (2006).
 - [4] National Research Council, *Developing Assessments for the next Generation Science Standards* (National Academies Press, Washington, DC, 2014).
 - [5] T. Lord and S. Baviskar, Moving students from information recitation to information understanding: Exploiting Bloom's taxonomy in creating science questions, *J. Coll. Sci. Teach.* **36**, 40 (2007), http://www.nsta.org/store/product_detail.aspx?id=10.2505/4/jcst07_036_05_40.
 - [6] J. L. Momsen, T. M. Long, S. A. Wyse, and D. Ebert-May, Just the facts? Introductory undergraduate biology courses focus on low-level cognitive skills, *CBE-Life Sci. Educ.* **9**, 379 (2010).
 - [7] P. Nichols and B. Sugrue, The lack of fidelity between cognitively complex constructs and conventional test development practice, *Educ. Meas. Issues Pract.* **18**, 18 (1999).
 - [8] L. B. Resnick and C. Zurawsky, Science education that makes sense, *Am. Educ. Res. Assoc. (AERA)* **5**, 1 (2007), <https://files.eric.ed.gov/fulltext/ED497647.pdf>.
 - [9] L. A. Shepard, The role of assessment in a learning culture, *Educ. Res.* **29**, 4 (2000).
 - [10] N. B. Songer, B. Kelcey, and A. W. Gotwals, How and when does complex reasoning occur? Empirically driven development of a learning progression focused on complex reasoning about biodiversity, *J. Res. Sci. Teach.* **46**, 610 (2009).
 - [11] D. Hestenes, M. Wells, and G. Swackhamer, Force concept inventory, *Phys. Teach.* **30**, 141 (1992).
 - [12] O. L. Liu, H.-S. Lee, and M. C. Linn, An investigation of explanation multiple choice items in science assessment, *Educ. Assess.* **16**, 164 (2011).
 - [13] S. A. Livingston, *Constructed-response test questions: Why we use them; How we score them* (Educ. Test. Serv. Princeton, NJ, 2009).
 - [14] D. F. Treagust, Development and use of diagnostic tests to evaluate students' misconceptions in science, *Int. J. Sci. Educ.* **10**, 159 (1988).
 - [15] D. C. Briggs, A. C. Alonzo, C. Schwab, and M. Wilson, Diagnostic assessment with ordered multiple-choice items, *Educ. Assess.* **11**, 33 (2006).
 - [16] H. O. Arslan, C. Cigdemoglu, and C. Moseley, A three-tier diagnostic test to assess pre-service teachers' misconceptions about global warming, greenhouse effect, ozone layer depletion, and acid rain, *Int. J. Sci. Educ.* **34**, 1667 (2012).
 - [17] I. S. Caleon and R. Subramaniam, Development and application of a three-tier diagnostic test to assess secondary students' understanding of waves, *Int. J. Sci. Educ.* **32**, 939 (2010).
 - [18] I. S. Caleon and R. Subramaniam, Do students know what they know and what they don't know? Using a four-tier diagnostic test to assess the nature of students' alternative conceptions, *Res. Sci. Educ.* **40**, 313 (2010).
 - [19] D. Gurcay and E. Gulbas, Development of three-tier heat, temperature and internal energy diagnostic test, *Res. Sci. Technol. Educ.* **33**, 197 (2015).
 - [20] D.-C. Yang and Y.-C. Lin, Assessing 10-to 11-year-old children's performance and misconceptions in number sense using a four-tier diagnostic test, *Educ. Res.* **57**, 368 (2015).
 - [21] P. Klein, A. Müller, and J. Kuhn, Assessment of representational competence in kinematics, *Phys. Rev. Phys. Educ. Res.* **13**, 010132 (2017).
 - [22] A. Hilton, G. Hilton, S. Dole, and M. Goos, Development and application of a two-tier diagnostic instrument to assess middle-years students' proportional reasoning, *Math. Educ. Res. J.* **25**, 523 (2013).
 - [23] W. L. Romine, T. D. Sadler, and A. T. Kinslow, Assessment of scientific literacy: Development and validation of the Quantitative Assessment of Socio-Scientific Reasoning (QuASSR), *J. Res. Sci. Teach.* **54**, 274 (2017).
 - [24] C.-Y. Tsui and D. Treagust, Evaluating secondary students' scientific reasoning in genetics using a two-tier diagnostic instrument, *Int. J. Sci. Educ.* **32**, 1073 (2010).
 - [25] C.-C. Tsai and C. Chou, Diagnosing students' alternative conceptions in science, *J. Comput. Assist. Learn.* **18**, 157 (2002).

- [26] A. L. Chandrasegaran, D. F. Treagust, and M. Mocerino, The development of a two-tier multiple-choice diagnostic instrument for evaluating secondary school students' ability to describe and explain chemical reactions using multiple levels of representation, *Chem. Educ. Res. Pract.* **8**, 293 (2007).
- [27] H.-E. Chu, D. F. Treagust, and A. L. Chandrasegaran, A stratified study of students' understanding of basic optics concepts in different contexts using two-tier multiple-choice items, *Res. Sci. Technol. Educ.* **27**, 253 (2009).
- [28] K. S. Taber and K. C. D. Tan, The insidious nature of 'hard-core' alternative conceptions: Implications for the constructivist research programme of patterns in high school students' and pre-service teachers' thinking about ionisation energy, *Int. J. Sci. Educ.* **33**, 259 (2011).
- [29] D. K.-C. Tan and D. F. Treagust, Evaluating students' understanding of chemical bonding, *Sch. Sci. Rev.* **81**, 75 (1999), <https://www.ase.org.uk/journals/school-science-review/1999/09/294/1186/SSR294Sept1999p75.pdf>.
- [30] D. F. Treagust and A. L. Chandrasegaran, The Taiwan national science concept learning study in an international perspective, *Int. J. Sci. Educ.* **29**, 391 (2007).
- [31] J.-R. Wang, Development and validation of a two-tier instrument to examine understanding of internal transport in plants and the human circulatory system, *Int. J. Sci. Math. Educ.* **2**, 131 (2004).
- [32] A. E. Lawson, The development and validation of a classroom test of formal reasoning, *J. Res. Sci. Teach.* **15**, 11 (1978).
- [33] V. P. Coletta and J. A. Phillips, Interpreting FCI scores: Normalized gain, preinstruction scores, and scientific reasoning ability, *Am. J. Phys.* **73**, 1172 (2005).
- [34] B. J. Franklin, Ph.D. thesis, The development, validation, and application of a two-tier diagnostic instrument to detect misconceptions in the areas of force, heat, light and electricity, The Louisiana State University and Agricultural & Mechanical College, 1992.
- [35] C.-C. Chen, H.-S. Lin, and M.-L. Lin, Developing a two-tier diagnostic instrument to assess high school students' understanding The formation of image by plane mirror, *Proc.-Natl. Sci. Counc. Repub. CHINA PART Math. Sci. Technol. Educ.* **12**, 106 (2002).
- [36] T. Fetherstonhaugh and D. F. Treagust, Students' understanding of light and its properties: Teaching to engender conceptual change, *Sci. Educ.* **76**, 653 (1992).
- [37] C. Kamcharean and P. Wattanakasiwich, Development and application of thermodynamics diagnostic test to survey students' understanding in thermal physics, *Int. J. Innov. Sci. Math. Educ. Former. CAL-Laborate Int.* **24**, 14 (2016), <https://openjournals.library.sydney.edu.au/index.php/CAL/article/view/8710>.
- [38] G. W. Fulmer, H.-E. Chu, D. F. Treagust, and K. Neumann, Is it harder to know or to reason? Analyzing two-tier science assessment items using the Rasch measurement model, *Asia-Pac. Sci. Educ.* **1**, 1 (2015).
- [39] B. K. Bayrak, Using two-tier test to identify primary students' conceptual understanding and alternative conceptions in acid base, *Mevlana Int. J. Educ.* **3**, 19 (2013), <https://journals.indexcopernicus.com/search/article?articleId=365358>.
- [40] H.-P. Chang, J.-Y. Chen, C.-J. Guo, C.-C. Chen, C.-Y. Chang, S.-H. Lin, W.-J. Su, K.-D. Lain, S.-Y. Hsu, J.-L. Lin *et al.*, Investigating primary and secondary students' learning of physics concepts in Taiwan, *Int. J. Sci. Educ.* **29**, 465 (2007).
- [41] C. Chou, P.-S. Chan, and H.-C. Wu, Using a two-tier test to assess students' understanding and alternative conceptions of cyber copyright laws, *Br. J. Educ. Technol.* **38**, 1072 (2007).
- [42] K. C. D. Tan, N. K. Goh, L. S. Chia, and D. F. Treagust, Development and application of a two-tier multiple choice diagnostic instrument to assess high school students' understanding of inorganic chemistry qualitative analysis, *J. Res. Sci. Teach.* **39**, 283 (2002).
- [43] A. L. Odom and L. H. Barrow, Development and application of a two-tier diagnostic test measuring college biology students' understanding of diffusion and osmosis after a course of instruction, *J. Res. Sci. Teach.* **32**, 45 (1995).
- [44] H. P. Tam, M. Wu, D. C. H. Lau, and M. M. C. Mok, Using user-defined fit statistic to analyze two-tier items in mathematics. in *Self-directed Learning Oriented Assessments in the Asia-Pacific* (Springer, New York, 2012), pp. 223–233.
- [45] C. Tüysüz, Development of two-tier diagnostic instrument and assess students' understanding in chemistry, *Sci. Res. Essays* **4**, 626 (2009), <http://www.academicjournals.org/journal/SRE/article-abstract/0B2C90A18467>.
- [46] S.-W. Lin, Development and application of a two-tier diagnostic test for high school students' understanding of flowering plant growth and development, *Int. J. Sci. Math. Educ.* **2**, 175 (2004).
- [47] J. Barbera, A psychometric analysis of the chemical concepts inventory, *J. Chem. Educ.* **90**, 546 (2013).
- [48] L. Bao, T. Cai, K. Koenig, K. Fang, J. Han, J. Wang, Q. Liu, L. Ding, L. Cui, Y. Luo *et al.*, Learning and scientific reasoning, *Science* **323**, 586 (2009).
- [49] L. Ding, Progression trend of scientific reasoning from elementary school to university: A large-scale cross-grade survey among Chinese students, *Int. J. Sci. Math. Educ.* **1** (2017).
- [50] J. Han, Ph.D. Dissertation, Scientific reasoning: Research, development, and assessment, The Ohio State University, 2013.
- [51] K. Koenig, M. Schen, and L. Bao, Explicitly targeting pre-service teacher scientific reasoning abilities and understanding of nature of science through an introductory science course, *Sci. Educ.* **21**, 1 (2012), <http://static.nsta.org/connections/college/201307Koenig.pdf>.
- [52] P. Nieminen, A. Savinainen, and J. Viiri, Relations between representational consistency, conceptual understanding of the force concept, and scientific reasoning, *Phys. Rev. ST Phys. Educ. Res.* **8**, 010123 (2012).
- [53] G. N. Masters, A Rasch model for partial credit scoring, *Psychometrika* **47**, 149 (1982).
- [54] E. Muraki, A generalized partial credit model: Application of an EM algorithm, *Appl. Psychol. Meas.* **16**, 159 (1992).
- [55] F. Samejima, Estimation of latent ability using a response pattern of graded scores, *Psychom. Monogr. Suppl.* (1969).

- [56] J.-X. Yao, Y.-Y. Guo, and K. Neumann, Towards a hypothetical learning progression of scientific explanation, *Asia-Pac. Sci. Educ.* **2**, 4 (2016).
- [57] J.-X. Yao and Y.-Y. Guo, Validity evidence for a learning progression of scientific explanation, *J. Res. Sci. Teach.* **55**, 299 (2017).
- [58] J. H. McDonald, *Handbook of Biological Statistics*, 3rd ed. (Sparky House Publishing Baltimore, MD, 2014).
- [59] S. S. Mangiafico, *Chi-square test of independence in Handbook of Biological Statistics*, 3rd ed. (Sparky House Publishing, Baltimore, Maryland, 2014), pp. 223–233.
- [60] L. Ding, X. Wei, and K. Molloy, Does higher education improve student scientific reasoning skills?, *Int. J. Sci. Math. Educ.* **14**, 619 (2016).
- [61] T. Bond and C. M. Fox, *Applying the Rasch Model: Fundamental Measurement in the Human Sciences* (Lawrence Erlbaum, Mahwah, NJ, 2007).
- [62] T. Kiefer, A. Robitzsch, and M. Wu, Test analysis modules, Comput. Softw. Man. Retrieved <http://CRAN.R-project.org/package=TAM> R PACKAGE VERSION 10-2 (2014).
- [63] J. M. Linacre, Winsteps (Version 3.70. 0) [Computer Software] (2010).
- [64] R. F. DeVellis, *Scale Development: Theory and Applications* (Sage Publications, Newbury Park, CA, 2012).
- [65] X. Liu, *Using learning progression to organize learning outcomes: Implications for assessment*, in *Making It Tangible—Learning Outcomes in Science Education*, edited by S. Bernholt, K. Neumann, and P. Nentwig (Waxmann, Münster, 2012), pp. 225–241.
- [66] T. Chai and R. R. Draxler, Root mean square error (RMSE) or mean absolute error (MAE)?—Arguments against avoiding RMSE in the literature, *Geosci. Model Dev.* **7**, 1247 (2014).
- [67] G. A. Morris, L. Branum-Martin, N. Harshman, S. D. Baker, E. Mazur, S. Dutta, T. Mzoughi, and V. McCauley, Testing the test: Item response curves and test quality, *Am. J. Phys.* **74**, 449 (2006).
- [68] G. A. Morris, N. Harshman, L. Branum-Martin, E. Mazur, T. Mzoughi, and S. D. Baker, An item response curves analysis of the Force Concept Inventory, *Am. J. Phys.* **80**, 825 (2012).
- [69] P. Kline, *A Handbook of Test Construction: Introduction to Psychometric Design* (Routledge, London, 1986), p. 143.
- [70] See Supplemental Material at <http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.14.020104> for additional analysis with populations S1, S3 and S4, and discussions of scoring model evaluation and selection.
- [71] R. J. Simes, An improved Bonferroni procedure for multiple tests of significance, *Biometrika* **73**, 751 (1986).